REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	27 Aug. 2001	Final Report: 1 Apr. 1998 - 31 Mar. 2001

4. TITLE AND SUBTITLE
Towards an Alternative for Antibodies: Construction and Characterization of a Large Combinatorial Library of Diverse Binding Proteins

5. FUNDING NUMBERS
N00014-95-1-0417

6. AUTHOR(S)
David Baker

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Washington
Department of Biochemistry
Seattle. Washington 98195

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217-5000

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Distribution Unlimited

20010910 112

13. ABSTRACT *(Maximum 200 words)*

The goal of this project is to create a large and very diverse population of binding proteins from which individual members can be selected on the basis of target specificity and be used as a substitute for antibodies in biosensor and other applications. Toward this end, phage display technology was used to create a collection of seventeen distinct combinatorial libraries in which the binding surfaces of several small, stable parent proteins were randomized. The completed collection contains approximately $4x10^9$ different protein variants. Because of the types of parent molecules chosen and because of the way the libraries were designed, the variants are expected to have more physical stability than antibody molecules and to be able to function in more severe types of environments. In a test of the general binding properties of the collection, the library pool was taken through four rounds of panning against 31 randomly chosen test compounds (20 proteins, 4 peptides and 7 small molecules). Apparent binding proteins were observed against 22/31 (71%) of the compounds tested (17 proteins, 2 peptides and 3 small organic molecules) indicating the library collection is a useful source of receptor proteins.

14. SUBJECT TERMS	15. NUMBER OF PAGES
combinatorial libraries, phage display, antibodies, biosensors	16. PRICE CODE: 4

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
unclassified	unclassified	unclassified	UL

FINAL REPORT

GRANT # N00014-95-1-0417

PRINCIPAL INVESTIGATOR: David Baker

INSTITUTION: University of Washington

GRANT TITLE: Towards An Alternative for Antibodies: Construction and Characterization of A Large Combinatorial Library of Diverse Binding Proteins

AWARD PERIOD: 1 April 1998 - 31 March 2001

OBJECTIVE: The overall goal of this project is to create a large and very diverse population of binding proteins from which individual members can be selected on the basis of target specificity and be used as a substitute for antibodies in biosensor and other applications. To accomplish this, we have synthesized a collection of combinatorial libraries in which the binding surfaces of several small, stable parent proteins are randomized. Because of the types of parent molecules chosen and because of the way the libraries are designed, the variants are expected to have more physical stability than antibody molecules and to be able to function in more severe types of environments.

APPROACH: Phage display technology is used to construct each library. The parent molecules selected for randomization are calmodulin, the c-src SH3 domain, the lck SH2 domain, and the immunoglobulin-binding domain from bacterial protein L. These proteins each have very different binding surface shapes and their natural targets cover a wide size range and include broad sections of large proteins, peptides, exposed loops of proteins and individual amino acid side chains. Library designs are customized for each protein and are based on the most current structural information to try to maximize the chance of producing variants with new target specificities relative to the parent molecule. In each case, residues responsible for maintaining the structural integrity of the protein fold were not changed. In the completed libraries, variant proteins are displayed on the surfaces of phagemid particles where they are folded and available for binding interactions. Individual variants are selected by multiple rounds of biopanning populations of library phagemids against immobilized target molecules. A single round of biopanning consists of bulk phase absorption of library phagemids to a test target followed by removal of weakly or non-specifically bound phagemids by washing, and then the elution, amplification and finally re-absorption of binders to the same test ligand to start a second round of biopanning. Variant proteins carried by phagemids obtained in such screens can be easily purified, cloned, expressed and adapted for use in specific applications.

ACCOMPLISHMENTS:

(1) Library pool completed.
Seventeen different libraries have been synthesized. Four libraries are variations of the c-src SH3 domain, two are from the lck SH2 domain, nine are from the protein L immunoglobulin-binding domain and two are from calmodulin. Complexities are between 10^7 and 10^8 protein

variants per individual library and there are a total of approximately $4x10^9$ different variants in the library pool.

(2) Recovery of proteins with new binding specificities from the library pool.

(a) Variant proteins which bind to a lambda-chain carrying immunoglobulin target that has no affinity for wild type protein L have been recovered from a library screen. The first variant characterized binds the new target with a disassociation constant of $4x10^{-7}$M.

(b) In order to evaluate the library pool and determine whether it is a practical source of receptors, we biopanned library phagemid populations against 31 different compounds. The collection of test compounds was chosen randomly and included 20 proteins, 4 peptides and 7 small molecules. None of the ligands showed measurable affinity for any of the library parent proteins. After four rounds of biopanning, we detected apparent binding proteins to 22 of the 31 ligands tested (71%). By "apparent binders" we the number of phagemids retained after washing were present at levels 10-100x greater than the background binding levels of phagemids to the negative controls and comparable to the highest levels of retention observed for the positive control targets. Of the 22 ligands able to select apparent binders out of the library pool, seventeen were proteins, two were peptides and three were small organic molecules. The original group of test ligands contained similar ratios of proteins, peptides and small molecules indicating there is no obvious target size binding preferences among the proteins in the library pool.

CONCLUSIONS: The library pool appears to be a good source of general binding proteins as it contained apparent binding proteins for more than 2/3 of the ligands tested. This is especially significant in that the ligands were chosen randomly and as a group, spanned a wide range of sizes, shapes and chemistries.

SIGNIFICANCE: Individual library variants will be suitable for many commercial and basic science applications. These include their use as receptors for biosensors, as purification reagents for affinity chromatography and as detection probes that can be used to identify specific proteins or individual post translational modifications within populations of proteins dispersed on Western blots or within cells processed for fluorescence microscopy.

AWARD INFORMATION:

NSF Young Investigator Award
Packard Foundation Fellowship
Protein Society Young Investigator Award
HHMI Assistant Investigatorship

PUBLICATIONS:

(1) Plaxco, K. W., Riddle, D. S., Grantcharova, V. and Baker, D. (1998). Simplified proteins: minimalist solutions to the 'protein folding problem'. Current Opinion in Structural Biology, 8: 80-85

(2) Grantcharova, V., Riddle, D. S., Santiago, J. V. and Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized

transition state for folding of the src SH3 domain. Nature Structural Biology, 5: 714-720

(3) Yi, Q., Bystroff, C., Rajagopal, P., Klevit, R. E. and Baker, D. (1998). Prediction and structural characterization of an independently folding substructure in the src SH3 domain. Journal of Molecular Biology, 283: 293-299.

(4) Riddle, D.S., Grantcharova, V.P., Santiago, J.V., Alm, E., Ruczinski, I. and Baker, D. (1999). Experiment and theory highlight role of native state topology in SH3 folding . Nature Structural Biology, 6: 1016-1024

(5) Grantcharova, V.P., Riddle, D.S. and Baker, D. (2000). Long-Range Order in the Src SH3 Folding Transition State. PNAS 97: 7084-7089.

(6) Yi, Q., Scalley-Kim, M.L., Alm, E.J. and Baker, D. (2000). NMR Characterization of Residual Structure in the Denatured State of Protein L. Journal of Molecular Biology, 299: 1341-1351.

(7) Minard, P., Scalley-Kim, M., Watters, A. and Baker, D. (2001). A "Loop Entropy Reduction" Phage-Display Selection For Folded Amino Acid Sequences. Protein Science, 10: 129-134.

(8) Grantcharova, V.P., and Baker, D. (2001). Circularization Changes the Folding Transition State of the Src SH3 Domain. Journal of Molecular Biology, 306: 555-163.

(9) Grantcharova, V.P., Alm, E., Baker, D., and Horwich, A. (2001). Mechanisms of Protein Folding. Current Opinion in Structural Biology, 11: 70-82

INVITED PRESENTATIONS:

(1) Scripts Research Institute. Spring, 1998
(2) John's Hopkins Folding Meeting. Spring, 1998
(3) Biopolymers Gordon Conference. Newport, RI. Summer 1998
(4) Centre European de Calcul Atomique et Moleculaire workshop. Turin, Italy. Summer, 1998
(5) American Chemical Society, Protein Structure Prediction Symposium, Boston, MA. Summer 1999.
(6) FASEB meeting on Protein Folding in the Cell. Vermont. Summer 1998.
(7) Departments of Molecular Biology and Chemistry, Princeton University. Fall 1998.
(8) 3rd Meeting for Critical Assessment of Protein Structure Prediction. Asilomar, CA. Fall 1998.
(9) IMS International Workshop on Protein Stability and Folding. Okazaki, Japan. Winter, 1999.
(10) Program in Mathematics and Molecular Biology. Santa Fe, NM. Winter, 1999.
(11) Department of Structural Biology, Stanford University. Winter, 1999.
(12) Bioinformatics '99. Lund, Sweden. Spring, 1999.
(13) Chemistry and Biology Departments, California Institue of Technology. Spring, 1999.

(14) MCB and Chemistry Departments, University of California, Berkeley. Spring, 1999.
(15) Institue of Molecular Biology, University of Oregon. Spring, 1999.
(16) Merck Evening Lecture Series, University of California, Berkeley. Spring, 19999.
(17) Symposium on Protein Sequence-Structure and Function. Dept. of Pharmaceutical Chemistry, University of California, San Francisco. Spring, 1999.
(18) American Physical Society Protein Folding Workshop. Atlanta, Spring, 1999.
(19) University of Toronto. Spring, 1999.
(20) Chemistry Department, University of Montreal. Spring, 1999.
(21) International Scientific Congress on Protein Folding (Belgium). Summer, 1999. (Member of Scientific organizing committee).
(22) Protein Society Symposium on Protein Structure Prediction. Summer, 1999.
(23) Department of Biochemistry, Northwestern University. Fall, 1999.
(24) Department of Biochemistry and Biophysics, Washington University School of Medicine. Fall 1999.
(25) Second Georgia Tech International Conference on Bioinformatics: In Silico Biology: Sequence & Structure & Function. Atlanta, Fall 1999
(26) ABRF 2000. "From Singular to Global Analyses of Biological Systems". Session chair and speaker. Bellevue, WA. Winter 2000.
(27) Symposium on "Computational Structural Biology in the Proteome Era". Tokyo, Japan. Winter 2000.
(28) Department of Cell Biology. University of Alabama, Birmingham. Winter 2000.
(29) Department of Biochemistry, Cell and Molecular Biology. University of Tennessee, Knoxville, TN. Winter 2000.
(30) Bioinformatics 2000. Elsinore, Denmark. Spring 2000.
(31) Protein Society Summer 2000
(32) American Physical Society Winter 2001
(33) American Chemical Society Spring 2001
(34) Hopkins Folding meeting Spring 2001
(35) University of Texas Structural Biology Symposium Spring 2001
(36) Proteins Gordon conference Summer 2001
(37) University of British Columbia Winter 2001
(38) UT Southwestern Winter 2001
(39) UCSF Winter 2001

Simplified proteins: minimalist solutions to the 'protein folding problem'

Kevin W Plaxco*, David S Riddle†, Viara Grantcharova‡ and David Baker#

Recent research has suggested that stable, native proteins may be encoded by simple sequences of fewer than the full set of 20 proteogenic amino acids. Studies of the ability of simple amino acid sequences to encode stable, topologically complex, native conformations and to fold to these conformations in a biologically relevant time frame have provided insights into the sequence determinants of protein structure and folding kinetics. They may also have important implications for protein design and for theories of the origins of protein synthesis itself.

Addresses
Department of Biochemistry, Box 357350, University of Washington, Seattle, WA 98195, USA
*e-mail: kwp@elina.bchem.washington.edu
†e-mail: riddle@u.washington.edu
‡e-mail: grantch@u.washington.edu
#e-mail: baker@ben.bchem.washington.edu

Current Opinion in Structural Biology 1998, 8:80–85

http://biomednet.com/elecref/0959440X00800080

Abbreviations
CD circular dichroism
SH3 Src homology 3

Introduction

Nature has solved the 'protein folding problem' countless times, generating thousands of families of rapidly folding, stable proteins with unique native conformations. The vast majority of these proteins are encoded by complex patterns of the 20 proteogenic amino acids [1–3]. This pattern and compositional complexity has proven to be a significant hurdle to theorists, experimentalists and engineers attempting to explain or reproduce the primary features of proteins. But is the full complexity of naturally occurring sequences required to encode their unique native structures? And, if not, what might 'simplified' sequences teach us about how a linear string of amino acids encodes a complex three-dimensional structure and can rapidly discriminate between this structure and the astronomically large number of conformations accessible to the unfolded state?

In this paper, we will review recent experiments aimed at designing or selecting for significantly simplified amino acid sequences capable of forming stable, native or native-like proteins. In discussing some of the insights into the sequence determinants of structure and folding kinetics gathered from these studies, we will place particular emphasis on recent investigations of simplified hydrophobic cores and fully simplified proteins, as well as on the ability of simplified proteins to rapidly fold to their native structures.

Simplified core packing

The tight packing observed in the interiors of proteins has led to the suggestion that the complementary shapes and sizes of core residues define a protein's fold in a fashion analogous to the manner in which the shapes of individual jigsaw pieces determine the overall layout of the finished puzzle (discussed in [4,5••,6–10]). Experimental tests of this hypothesis, however, have repeatedly demonstrated that dramatic core changes can be accommodated without significantly disrupting native structure [5••,6,11,12]. But although these studies indicate that the rules of core packing are fairly flexible, the variant proteins investigated all maintain the high degree of core complexity characteristic of naturally occurring proteins. Thus the question remains: is it possible for a simple amino acid sequence, presumably lacking highly complementary side-chain interactions, to encode a well-packed hydrophobic core? Recent evidence suggests that it is.

An earlier series of core simplification attempts (reviewed in [13]) is illustrative of the approach. Regan and co-workers [14,15] have produced variants of the RNA-binding protein Rop with highly simplified hydrophobic cores. Rop is a semi-regular dimer of two-helix monomers packed to form a four-helix bundle. Like coiled-coil proteins, the sequence of Rop is an array of heptad repeats (*abcdefgabcdefg*) in which positions *a* and *d* contribute to the hydrophobic core. In the wild-type protein, six of 16 *a* sites and eight of 16 *d* sites are alanine and leucine respectively. A variant of Rop with all-alanine *a* sites and all-leucine *d* sites is significantly more thermostable than the wild-type protein and folds into a fully native protein [14,15]. An alternating sequence of large and small hydrophobic residues may be critical for the formation of native core packing: the over-packed all-leucine variant forms a very stable molten globule and the under-packed all-alanine variant remains unstructured (Table 1). This suggests that, perhaps because of constraints associated with the symmetrically packed helical structure of Rop, a two-letter core alphabet and limited pattern complexity are required to encode a native structure.

The simplification of the core of Rop was based on, and perhaps somewhat limited by, the regular structure of four-helix bundles, but studies of globular proteins indicate that they may be even more amenable to simplification.

Table 1

Some simplified proteins and the sequences from which they were derived.

Name	ΔG_u (kcal/mol)	k_f (s^{-1})	Simplification scheme	References
WT Rop	7.7	0.013		[14,15,37••]
Ala2Leu2–8	7.5	7.9	Alternating leucine-alanine core	
Ala4–8	<0		All alanine core	
Leu4–8	30		All leucine core	
WT T4 lysozyme				[16••]
7 met T4 lysozyme	(–5.0)*		7 of 10 core positions: methionine	
10 met T4 lysozyme	(–7.3)*		10 of 10 core positions: methionine	
WT Cro repressor	2.0†			[17]
M-Cro	0.4		11 of 13 core positions: leucine	
α1	11.4		Simple helical homotetramer (GDLK)	[22–24]
α4	15.4		Plus helix cap and turns (GDLKPR)	
α_4-His_4	10.3		Plus metal binding site (GDLKPRH)	
WT SrcSH3	3.7	57		[26••]
FP1	3.0	93	40 of 45 non-active site residues (GEIKA)	
FP2	1.7	57	39 of 45 non-active site residues (GEIKA)	

*Absolute unfolding free energies are not available. Unfolding free energies ($\Delta\Delta G_u$) relative to wild-type are reported. †Stability of point mutant from which simplified variants were derived. FP, full protein; WT, wild-type.

Wild-type T4 lysozyme consists of two domains, the larger of which contains a hydrophobic core of approximately 10 residues. Matthews and co-workers [16••] have recently characterized variants with seven and 10 of these residues replaced by methionine. Surprisingly, although somewhat destabilized (Table 1), even the 10-methionine variant folds to a partially active and thus presumably very native-like conformation. Crystallographic analysis of the seven-methionine variant indicates that the core is so well packed that the total volume of the new core methionines is slightly less than would be predicted from the density of well-packed methionine crystals [4]. Although the average backbone deviation between the wild type and seven-methionine variant is only 0.2 Å, changes of up to 1.0 Å are observed at several backbone positions. It is via these small conformational rearrangements that the simplified variant is able to accommodate such dramatic compositional changes and generate this well-packed core.

Studies of simplified core variants of the phage 434 Cro protein suggest that the flexible, unbranched side chain of methionine is not required to create highly simplified hydrophobic cores [17]. One variant, with 11 of 13 core residues replaced with leucine, exhibits the folding cooperativity and NMR dispersion expected of a fully folded protein, although the variant is significantly destabilized (Table 1). These examples of well-packed hydrophobic cores constructed from extremely simple amino acid sequences suggest that the packing of core residues need not mimic the precise spatial complementarity of the pieces of a jigsaw puzzle in order to encode a unique, native conformation.

Simplified proteins

If hydrophobic cores can be built from one or two residue types and very simple sequence patterns, what are the minimum size amino acid alphabet and the simplest sequence patterns that can encode entire proteins? Several important 'alanine minimization' studies have confirmed that much of the sequence of naturally occurring globular proteins is redundant and can be replaced by alanine provided a 'scaffold' of residues is maintained that encodes a hydrophobic core and defines important secondary structural elements and tertiary contacts [18,19]. Hecht and co-workers [5•,20,21•] have demonstrated that four-helix bundle proteins can be generated from highly divergent sequences of 11 of the proteogenic amino acids, as long as the correct binary pattern of hydrophobic and hydrophilic residues is maintained. Although the non-native structures of the all-alanine and all-leucine Rop core variants suggest that many of these sequences do not encode fully native proteins, the extremely high recovery (60%) of soluble, protease-resistant and potentially native-like species reported suggests that, with a well-chosen reduced alphabet and the correct patterning of hydrophobic and hydrophilic residues, even compositionally simple sequences are likely to be able to encode folded proteins.

DeGrado and co-workers [22–24] have been exploring this issue through a series of elegant studies directed towards the *de novo* design of four-helix bundle proteins. Their earlier efforts, aimed at simple, multimeric proteins built entirely of glycine, glutamate, leucine and lysine culminated in the production of short helical segments capable of forming four-helix bundle like tetramers in

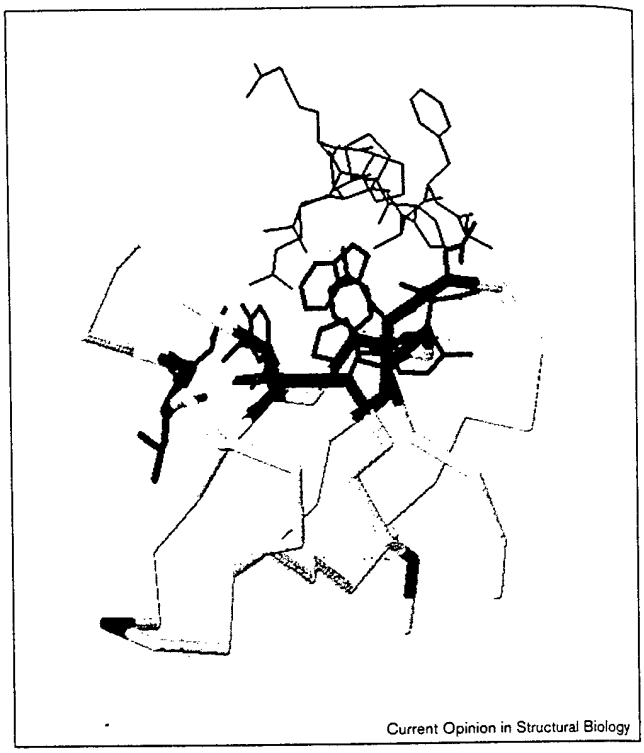
solution [22]. The addition of proline residues to break the helices and arginine residues to create interhelix loops resulted in the creation of a monomeric four-helix bundle [22,23]; however, despite the relatively high stability of the four-helix bundle (Table 1), it exhibits the rather poor hydrogen exchange protection and non-unique core packing [24] characteristic of the over-packed, all-leucine core variant of Rop. The formation of truly native four-helix bundles required the addition of several histidine residues to form a metal-binding site [24], although comparison with the Rop structure suggests that the addition of a limited degree of core complexity might also produce a native structure.

Simplified helical bundle proteins take advantage of the simple, periodic amino acid patterns of amphipathic helices. Work in our laboratory, however, has suggested that it is also possible to build complex, nonregular topologies from reduced amino acid alphabets. Using a phage display approach (see [25] and references therein), we have selected for significantly simplified sequences that fold into the structure of the topologically complex, predominantly β-sheet SH3 (Src homology 3) domain. And how few residue types does it take to make this topologically complex structure? Attempts to select for proteins comprising a three-residue alphabet (lysine, isoleucine and glutamate) were relatively unsuccessful as numerous alanine and glycine residues were maintained in the few folded proteins recovered. These two residues are probably required because of the need for a small nonpolar residue to pack with the large, β-branched isoleucine and because of the propensity of glycine to form tight turns. The inclusion of these two residues into the reduced alphabet led to the recovery of several highly simplified sequences that adopt the SH3 fold. The two simplest variants recovered to date are 68% and 70% composed of this five-residue set (89% and 90% respectively of residues outside the active site; Figure 1) and yet their stability (Table 1), activity, NMR and circular dichroism (CD) spectra [26••] are those of fully native proteins. Although these variants are almost as compositionally simplified as the helical proteins described above, they are encoded by highly complex, nonrepetitive amino acid sequences. This is consistent with the observation that naturally occurring β-sheet proteins lack the strong patterning of hydrophobic and hydrophilic residues apparent in their α-helical counterparts [2]. These results indicate that a five-residue alphabet may be sufficient to encode all of the functions required to generate even topologically complex proteins.

The simplest proteins

If four-helix bundles and globular β-sheet proteins can be encoded by simple amino acid sequences, what fraction of simple, random sequences might encode native proteins? Recent work conducted in our laboratory and by Sauer and co-workers [27–29] indicates that a surprisingly large fraction of even the very simplest sequences fold, although perhaps not to fully native proteins, to at least protein-like structures. Sauer and co-workers [27–29] have investigated simple, 80-residue proteins made up almost entirely of glutamine, leucine and arginine. Approximately 1% of these randomly generated sequences encode proteins that are sufficiently protease resistant to be recovered from the *Escherichia coli* expression library used in the study, of which approximately two-thirds display cooperative thermal denaturation transitions and CD signatures characteristic of folded, helical proteins. Many of these proteins also form apparently well-defined multimeric complexes similar to those formed by naturally occurring proteins. Hydrogen-exchange experiments, however, indicate that although the recovered proteins are native-like, they exhibit the reduced levels of hydrogen-exchange protection characteristic of the molten globule state. The fraction of such sequences that encode fully native proteins remains to be determined.

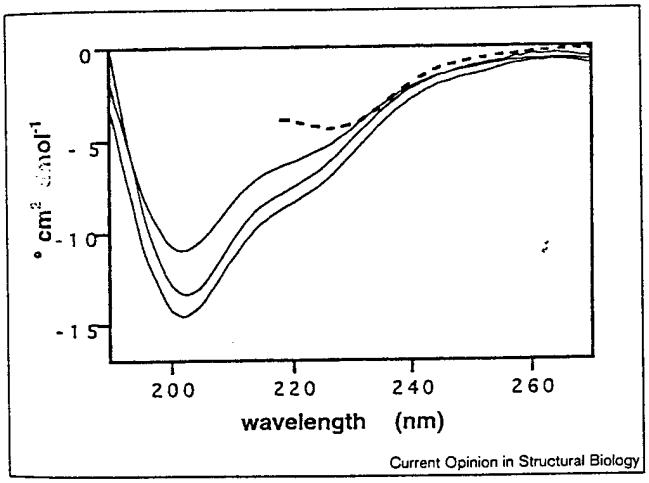
Figure 1



A model of the structure of a highly simplified SH3 domain. 68% of FP2 is composed of just five residue types (isoleucine, lysine, aspartate, glycine and alanine), which are shown as a grey backbone ribbon. The active site and recognition peptide residues are rendered in black with side chains, and residues not simplified in FP2 shown as black tubes. 90% of residues outside the binding cleft are composed of the reduced amino acid alphabet. FP2 is the most simplified protein recovered so far from simple sequence SH3 domain libraries but may not represent the simplest possible SH3 sequence. The observation of other variants simplified at several of the positions not simplified in FP2 suggests that these positions do not necessarily encode unique structural information. The structure, stability and folding kinetics of FP2 suggest that the full complexity of naturally occurring sequences is not required to encode rapidly folding, topologically complex, native proteins [26••]. FP, full protein.

Our laboratory has studied the properties of a chemically synthesized 'sequence space soup' of heteropolymers composed of leucine (40%), ornithine (30%) and glutamate (30%). Populations of random sequences with this bulk composition and of homogeneous molecular weight were generated. They were readily soluble in aqueous buffer and exhibited considerable helical structure that was disrupted by guanidine hydrochloride denaturation (Figure 2). Gel filtration and small angle X-ray scattering experiments indicate that the majority of these random peptides form small oligomers (V Grantcharova, D Baker, unpublished data). Overall, the properties of the population were similar to that of those individual sequences characterized by Sauer and co-workers [27–29], suggesting that helicity and the tendency to form small oligomers are properties frequently encoded by this class of simple sequences.

Figure 2



Sequence space soups comprising 23-, 46- or 70-residue random polymers of leucine, glutamate and ornithine (solid lines from top down) fold to form α-helical structures as indicated by these CD spectra. Because these amino acids have the same molecular weight, all of the peptides in each population are of equal masses. The helical structure of the 70-residue peptides is lost in the presence of guanidine hydrochloride (dotted line). Despite the constant molecular mass, the 70-mer population exhibits hydrodynamic behavior consistent with the formation of small oligomers (data not shown).

The folding kinetics of simplified sequences

In 1936, Mirsky and Pauling [30] correctly surmised that the renaturation of a denatured protein requires that a single native structure be distinguished from among approximately 10^{20} possible unfolded conformations. In 1961, Anfinsen [31] demonstrated that all of the information required to rapidly perform this discrimination is encoded by a protein's primary sequence. More than thirty-five years later, investigations of the mechanisms by which this 'kinetic half of the protein folding problem' is resolved remain an area of active research. Recent studies of the folding kinetics of highly simplified proteins have attempted to define the sequence determinants of this process.

A fundamental aspect of the folding of proteins is that an extended and highly disordered polymer chain must collapse to form a compact, globular protein. It has been postulated that a largely random collapse process, driven by the exclusion of hydrophobic residues from the solvent, is followed by the reorganization of this compact state to form the native conformation (reviewed in [32,33,34•]). If this theory is correct, hydrophobic core simplification might be expected to disrupt collapse efficiency and reduce otherwise optimized folding rates. Moreover, if the diffusive rearrangements of a collapsed intermediate is a rate-limiting folding step, then core sequence redundancy might slow folding as degenerate conformations with near-native packing trap the rearrangement process [35•,36].

Recent studies of the folding kinetics of simplified core Rop variants do not support a role for precise chain packing in directing the folding process. All of the simplified core variants of Rop fold more rapidly than does the wild-type protein (Table 1)—the alternating alanine and leucine core variant an amazing 610 times more rapidly [37••]. Although this acceleration may be due to the elimination of hydrophilic core residues [37••,38], that the refolding rates of the simplified variants are not decelerated indicates that highly unique, jigsaw-like core packing might not be a fundamental requirement of rapidly folding proteins. A small protein like Rop, however, might not accurately model the folding processes of larger proteins, which are thought to fold in a two-step process that is much more dependent on the rate of core reorganization in collapsed intermediate species [32,33,34•]. We thus eagerly await the results of studies of the refolding of simplified core variants of larger proteins.

If specific core packing is not required to generate rapidly folding proteins, then what are the determinants of protein folding kinetics? Recent investigations of the folding rate of simplified SH3 variants suggest instead that folding rates may be largely determined by the interactions that stabilize the native state and might not depend on specific, conserved folding pathways [39]. Despite dramatic changes in amino acid sequence and overall amino acid composition, the two most highly simplified SH3 variants characterized to date refold as fast as or faster than the wild-type SH3 sequence from which they were derived (Table 1) [26••]. Moreover, the simplified variants and the wild-type sequence maintain qualitatively very similar folding kinetics: all fold in simple, two-state processes via relatively similar transition state conformations. That these variants fold rapidly despite the absence of any obvious selection for this characteristic suggests that rapid folding is more a feature of a stable, cooperative native fold, which was a selection criterion, than of specific, conserved folding pathways. The rapid folding of these simple variants also suggests

that the compositional complexity found in naturally occurring proteins may not be required to encode their rapid folding.

Simple proteins and the origins of protein synthesis

Protein synthesis is a complex process involving numerous tRNAs, acyl-tRNA synthetases and many other critical components. How might such a complex system have arisen spontaneously? Despite a long history of investigation (see, for example, [40–44]) no answer to this question appears forthcoming. The ability of reduced alphabet proteins to fold rapidly to stable, native structures suggests, however, that the full complexity seen in contemporary protein synthesis might not have been required to generate primordial proteins capable of providing a selective advantage [45].

Conclusions

Although significant advances have been made towards the goal of constructing highly compositionally simplified proteins, it is notable that native proteins composed of less than seven amino acid types have not yet been demonstrated. This is no more simple than the simplest structured, naturally occurring proteins. (The single-helix, 48-residue yellowtail flounder antifreeze protein, to our knowledge, holds the record at just seven residue types [46,47•].) Although this may suggest that more highly simplified native proteins cannot exist, the preceding studies have demonstrated that even smaller subsets of the proteogenic amino acids can encode each of the individual properties characteristic of naturally occurring proteins. Thus we are optimistic that future attempts to generate more significantly simplified proteins will prove fruitful. The demonstration of such proteins should facilitate *de novo* protein design efforts by indicating the minimal sequence elements required to encode native folds. These studies should also serve to filter the 'noise' out of naturally occurring protein sequences, and project into sharp contrast the major determinants of structure and folding.

Note added in proof

Stroud and co-workers [48••] have recently published the synthesis and structure of a *de novo* designed 108-residue protein composed of just seven residue types that folds into a fully native four-helix bundle.

Acknowledgements

The authors wish to thank W Baas, V Daggett, M Gross, L Regan, D Shortle, J Yang and our co-workers in the Baker group for their contributions to this review. We also thank Ron Zuckerman for the chemically synthesized sequence space soups.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Irback A, Peterson C, Potthast F: **Evidence for nonrandom hydrophobicity structures in protein chains.** *Proc Natl Acad Sci USA* 1996, **93**:9533-9538.

2. West MW, Hecht MH: **Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins.** *Protein Sci* 1995, **4**:2032-2039.

3. Wootton JC: **Sequences with 'unusual' amino acid compositions.** *Curr Opin Struct Biol* 1994, **4**:413-421.

4. Chothia C, Gerstein M: **How far can sequences diverge?** *Nature* 1997, **385**:579-581.

5. •• Beasley JR, Hecht MH: **Protein design: the choice of *de novo* sequences.** *J Biol Chem* 1997, **272**:2031-2034.
An excellent review of the sequence determinants of protein structure. The authors place particular emphasis on the role played by simple binary patterns in encoding the three-dimensional structure of the native state.

6. Sauer RT: **Protein folding from a combinatorial perspective.** *Fold Des* 1996, **1**:R27-R30.

7. Bromberg S, Dill KA: **Side-chain entropy and packing in proteins.** *Protein Sci* 1994, **3**:997-1009.

8. Behe MJ, Lattman EE, Rose GD: **The protein-folding problem: the native fold determines packing, but does packing determine the native fold?** *Proc Natl Acad Sci USA* 1991, **88**:4195-4199.

9. Ponder JW, Richards FM: **Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes.** *J Mol Biol* 1987, **193**:775-791.

10. Crick FHC: **The packing of α-helices: simple coiled-coils.** *Acta Crystallogr* 1953, **6**:689-697.

11. O'Brien R, Wynn R, Driscoll PC, Davis B, Plaxco KW, Sturtevant JM, Ladbury JE: **The adaptability of *Escherichia coli* thioredoxin to non-conservative amino acid substitutions.** *Protein Sci* 1997, **6**:1325-1332.

12. Axe DD, Foster NW, Fersht AR: **Active barnase variants with completely random hydrophobic cores.** *Proc Natl Acad Sci USA* 1996, **93**:5590-5594.

13. Clarke ND: **Sequence 'minimization': exploring the sequence landscape with simplified sequences.** *Curr Opin Biotechnol* 1995, **6**:467-472.

14. Munson M, O'Brien R, Sturtevant JM, Regan L: **Redesigning the hydrophobic core of a four-helix-bundle protein.** *Protein Sci* 1994, **3**:2015-2022.

15. Munson M, Balasubramanian S, Fleming KG, Nagi AD, O'Brien R, Sturtevant JM, Regan L: **What makes a protein? Hydrophobic core designs that specify stability and structural properties.** *Protein Sci* 1996, **5**:1584-1593.

16. •• Gassner NC, Baase WA, Matthews BW: **A test of the 'jigsaw puzzle' model for protein folding by multiple methionine substitutions within the core of T4 lysozyme.** *Proc Natl Acad Sci USA* 1996, **93**:12155-12158.
This crystallographic and thermodynamic study of T4 lysozyme variants with significantly simplified hydrophobic cores demonstrates clearly the extraordinary degree of core compositional simplicity that a globular protein can accommodate and poses a serious challenge to the 'jigsaw puzzle' model of core packing and protein structure.

17. Desjarlais JR, Handel TM: ***De novo* design of the hydrophobic cores of proteins.** *Protein Sci* 1995, **4**:2006-2018.

18. Michael SF, Kilfoil VJ, Schmidt MH, Amann BT, Berg JM: **Metal binding and folding properties of a minimalist Cys2His2 zinc finger peptide.** *Proc Natl Acad Sci USA* 1992, **8**:4796-4800.

19. Shang Z, Isaac VE, Li H, Patel L, Catron KM, Curran T, Montelione GT, Abate C: **Design of a 'minimAl' homeodomain: the N-terminal arm modulates DNA binding affinity and stabilizes**

homeodomain structure. *Proc Natl Acad Sci USA* 1994, **91**:8373-8377.

20. Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH: **Protein design by binary patterning of polar and nonpolar amino acids.** *Science* 1993, **265**:1680-1685.

21. • Roy S, Helmer KJ, Hecht MH: **Detecting native-like properties in combinatorial libraries of *de novo* proteins.** *Fold Des* 1997, **2**:89-92.

NMR is used in this study to investigate the nativeness of representative proteins derived from a library of random sequences that conform to a binary pattern thought to encode four-helix bundle proteins.

22. DeGrado WF, Wasserman ZR, Lear JD: **Protein design, a minimalist approach.** *Science* 1989, 243:622-628.

23. Regan L, DeGrado WF: **Characterization of a helical protein designed from first principles.** *Science* 1988, **241**:976-978.

24. Handel TM, Williams SA, DeGrado WF: **Metal ion-dependent modulation of the dynamics of a designed protein.** *Science* 1993, **261**:879-885.

25. Gu H, Yi Q, Bray ST, Riddle DS, Shiau AK, Baker D: **A phage display system for studying the sequence determinants of protein folding.** *Protein Sci* 1995, 4:1108-1117.

26. •• Riddle DS, Santiago JV, Bray ST, Doshi N, Grantcharova V, Yi Q, Baker D: **Functional rapidly folding proteins from simplified amino acid sequences.** *Nat Struct Biol* 1997, 4:805-809.

A phage display technique was used to select for significantly simplified proteins that adopt an SH3 domain fold. These studies suggest that a five-letter alphabet is the minimal compositional complexity required to generate this predominantly β-sheet fold. The rapid folding of the highly simplified variants recovered in this work (one folds 50% more rapidly than the wild-type sequence from which it was derived) suggests that sequence complexity of naturally occurring proteins is not required to generate rapidly folding species.

27. Cordes MHJ, Davidson AR, Sauer RT: **Sequence space, folding and protein design.** *Curr Opin Struct Biol* 1996, 6:3-10.

28. Davidson AR, Sauer RT: **Folded proteins occur frequently in libraries of random amino acid sequences.** *Proc Natl Acad Sci USA* 1994, **91**:2146-2150.

29. Davidson AR, Lumb KJ, Sauer RT: **Cooperatively folded proteins in random sequence libraries.** *Nat Struct Biol* 1995, 2:856-864.

30. Mirsky AK, Pauling L: **On the structure of native, denatured, and coagulated proteins.** *Proc Natl Acad Sci USA* 1936, **22**:439-447.

31. Anfinsen CB: **Principles that govern the folding of protein chains.** *Science* 1973, **181**:223-230.

32. Miranker AD, Dobson CM: **Collapse and cooperativity in protein folding.** *Curr Opin Struct Biol* 1996, 6:31-42.

33. Ptitsyn OB: **Structures of folding intermediates.** *Curr Opin Struct Biol* 1995, 5:74-78.

34. • Roder H, Colón W: **Kinetic role of early intermediates in protein folding.** *Curr Opin Struct Biol* 1997, 7:15-28.

An excellent review of collapse and cooperativity in protein folding and the role of collapse as a rate-limiting step in the folding process.

35. • Hinds DA, Levitt M: **From structure to sequence and back again.** *J Mol Biol* 1996, **258**:201-209.

This theoretical study explores some potential structural pitfalls faced by proteins composed of reduced amino acid alphabets.

36. Chan HS, Dill KA: **Comparing folding codes for proteins and polymers.** *Proteins* 1996, 24:335-344.

37. •• Munson M, Anderson KS, Regan L: **Speeding up protein folding: mutations that increase the rate at which Rop folds and unfolds by over four orders of magnitude.** *Fold Des* 1997, **2**:77-87.

This investigation into the folding rates of core-simplified variants of Rop suggests that the potential packing redundancy of simplified hydrophobic cores does not reduce protein folding rates. Indeed, all of the simplified variants investigated fold more rapidly than the wild-type sequence from which they were derived, potentially due to the removal of buried hydrophilic residues.

38. Waldburger CD, Jonsson T, Sauer RT: **Barriers to protein folding: formation of buried polar interactions is a slow step in acquisition of structure.** *Proc Natl Acad Sci USA* 1996, **93**:2629-2634.

39. Levinthal C: **Are there pathways for protein folding?** *J Chem Phys* 1968, **65**:44-45.

40. Crick FHC: **The origin of the genetic code.** *J Mol Biol* 1968, **38**:367-379.

41. Crick FHC, Orgel L: **Directed Panspermia.** *Icarus* 1973, **19**:341-346.

42. Wong JT: **A co-evolution theory of the genetic code.** *Proc Natl Acad Sci USA* 1975, **72**:1909-1912.

43. Kuhn H, Waser J: **On the origin of the genetic code.** *FEBS Lett* 1994, **352**:259-264.

44. Moller W, Janssen GM: **Transfer RNAs for primordial amino acids contain remnants of a primitive code at position 3 to 5.** *Biochimie* 1990, **72**:361-368.

45. Abkevich VI, Gutin AM, Shakhnovich EI: **How the first biopolymers could have evolved.** *Proc Natl Acad Sci USA* 1996, **93**:839-844.

46. Sicheri F, Yang DS: **Ice-binding structure and mechanism of an antifreeze protein from winter flounder.** *Nature* 1995, **375**:427-431.

47. • Davies PL, Sykes BD: **Antifreeze proteins.** *Curr Opin Struct Biol* 1997, 7:828-834.

A comprehensive review of the structures and activities of some of the simplest naturally occurring proteins.

48. •• Schafmeister CE, LaPorte SL, Miercke LJW, Stroud RM: **A designed four helix bundle protein with native-like structure.** *Nat Struct Biol* 1997, 4:1039-1042.

This crystallographic study of a *de novo* designed, seven-residue type, 108 residue, four-helix bundle clearly demonstrates that only limited compositional and pattern complexity are required to generate truly native proteins

Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain

Viara P. Grantcharova[1,2], David S. Riddle[1,2], Jed V. Santiago[1] and David Baker[1]

Experimental and theoretical studies on the folding of small proteins such as the chymotrypsin inhibitor 2 (CI-2) and the P22 Arc repressor suggest that the folding transition state is an expanded version of the native state with most interactions partially formed. Here we report that this picture does not hold generally: a hydrogen bond network involving two β-turns and an adjacent hydrophobic cluster appear to be formed in the folding transition state of the src SH3 domain, while the remainder of the polypeptide chain is largely unstructured. Comparison with data on other small proteins suggests that this structural polarization is a consequence of the topology of the SH3 domain fold. The non-uniform distribution of structure in the folding transition state provides a challenging test for computational models of the folding process.

A detailed understanding of how amino acid sequences determine protein three-dimensional structures requires the identification of residues and interactions that play critical roles in the folding process. In previous experiments directed at establishing the minimal sequence requirements for the folding of a small protein, the SH3 domain, a combinatorial library selection strategy was used to obtain functional SH3 domains composed primarily of I, K, E, A and G outside of the binding site[1]. The folding rates of these simplified SH3 variants were found to be very similar to that of the wild type protein, suggesting that the residues critical to folding kinetics must have been conserved in the selection. Here we investigate the roles of the conserved residues (Table 1) in the folding reaction by studying the consequences of alanine substitutions at these positions on the thermodynamics and kinetics of folding. Initially we focused on a non-local hydrogen bonding network involving Glu 30, Ser 47 and Thr 50 (Fig. 1) due to the strong selection pressure observed at these positions (Glu 30 is absolutely conserved and Ser 47 is frequently recovered even though it was not allowed in the mutagenesis strategy). The finding that these residues are important for the kinetics of folding prompted us to extend the analysis to residues throughout the src SH3 domain in order to obtain a more complete picture of the folding transition state.

Contributions to src SH3 stability

The structure of the src SH3 domain consists of two β-sheets orthoganally packed around a hydrophobic core[2,3]. Within the sheets, strands are joined by the RT, n-src and distal loops, while the crossovers between the two sheets occur at a diverging type II β-turn and a short 3_{10}-helix (Fig 1*a*). The residues conserved in the combinatorial library selection were located in the distal loop and diverging turn (Gly 29, Gly 51), the hydrogen bond network between them (Glu 30, Ser 47 and Thr 50; Fig. 1*b*), and the hydrophobic core (Phe 10, Leu 24, Ile 34, Ala 45, Ile 56). Substitutions in the hydrophobic core were most destabilizing, with mutations in the center of the core (Ala 45, Leu 32, Ile 56) having larger effects ($\Delta\Delta G_U$ from 1.8–3.6 kcal mol^{-1}) than mutations at the periphery (Leu 24, Phe 10, Val 61, Ile 34; $\Delta\Delta G_U$ from 0.6–1.8 kcal mol^{-1}). Mutations in the hydrogen bonding network also decreased ΔG_U significantly ($\Delta\Delta G_U$ from 1.8–2.5 kcal mol^{-1}), indicating the importance of these interactions in stabilizing the native state. Each of the mutated residues makes both local and non-local hydrogen bonds and thus the total energetic cost of the mutations is consistent with previous estimates of 1–2 kcal mol^{-1} per hydrogen bond[4]. Partially exposed aromatic residues lining the peptide binding pocket (Trp 42, Tyr 16, Tyr 55) can be viewed as extensions of the hydrophobic core and alanine substitutions at these positions decreased stability. On the other hand, mutation of the completely solvent exposed Tyr 60 to alanine increased ΔG_U, probably by destabilizing non-native conformations in which this residue is partially buried. The remaining mutations probe the integrity of the various loops. Glycine to alanine substitutions in the distal loop (G51) and the diverging turn (G29) decreased stability, while disrupting interactions in the n-src (G40, N36) and RT loops (D15, S18) either did not affect or slightly increased stability. The interpretation that the first two structural elements have more rigid structural requirements than the second two is consistent with our previous finding that amide protons in the distal loop hairpin and diverging turn are more protected from exchange than amide protons in the n-src and RT loops [5].

Kinetics of folding and unfolding

The kinetics of folding of the mutant proteins were characterized using stopped-flow fluorescence. The mutations were found to fall into three categories depending on whether the folding rate (k_f, Fig. 2*a*), the unfolding rate (k_u, Fig. 2*b*), or both were affected (Table 2). It is convenient to use simple transition state theory to interpret the kinetic data; computational models of folding for which the folding rate and the free energy difference between the unfolded state and the transition state ($\Delta G_{U-\ddagger}$) can be determined independently[6,7], suggest that the approximation rate $\equiv$ D exp ($-\Delta G_{U-\ddagger}/RT$) provides an excellent estimate of the folding rate (D is the frequency of transitions between related

[1]Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA. [2]Both of these authors contributed equally to this work.

Correspondence should be addressed to D.B. *email: baker@ben.bchem.washington.edu*

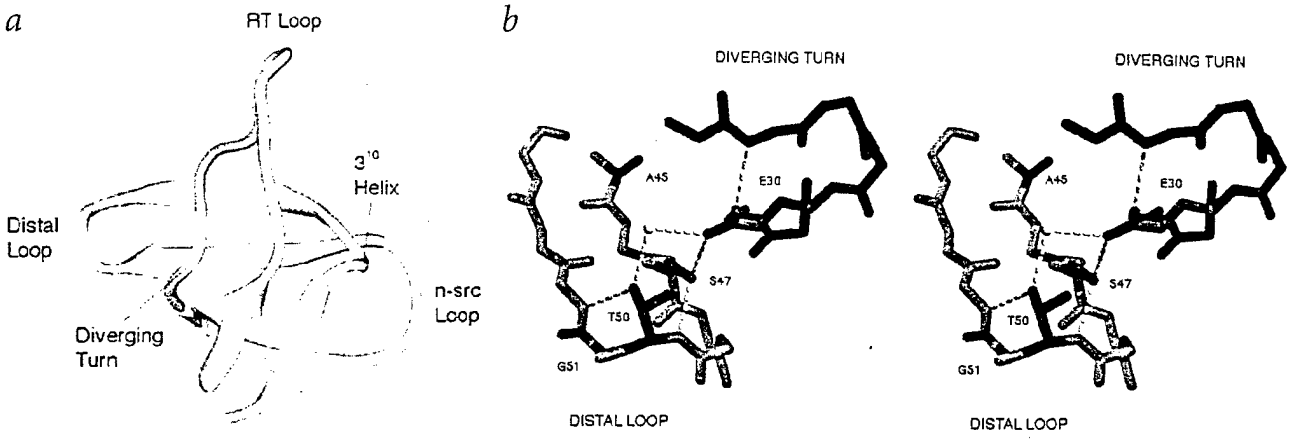


Fig. 1 ***a***, Ribbon diagram of the src SH3 domain crystal structure[2] with the loops and turns labeled. ***b***, Non-local hydrogen bond network connecting the distal loop (Ser 47 and Thr 50) and the diverging turn (Glu 30). Atomic coordinates were taken from the crystal structure of src SH3 domain within the context of the intact tyrosine kinase [2]. A coordinating water molecule may stabilize the interaction. Hydrogen bonding residues are shown in red; the two other residues in this region with high Φ_F values (Gly 51 and Ala 45) are shown in magenta. Both images were created with MidasPlus[38,39].

conformations). A simple but useful interpretation of the kinetic data based on this expression is that mutations which decrease k_f, but do not alter k_u disrupt interactions stabilizing both the transition and the native state while mutations which increase k_u, but do not change k_f disrupt interactions formed after the transition state[8]. A mutation will simultaneously decrease k_f and increase k_u in this model if some but not all of the interactions made by the residue in the native state are also made in the transition state. The structure of the folding transition state for the src SH3 domain deduced from the kinetic measurements using this model is presented below.

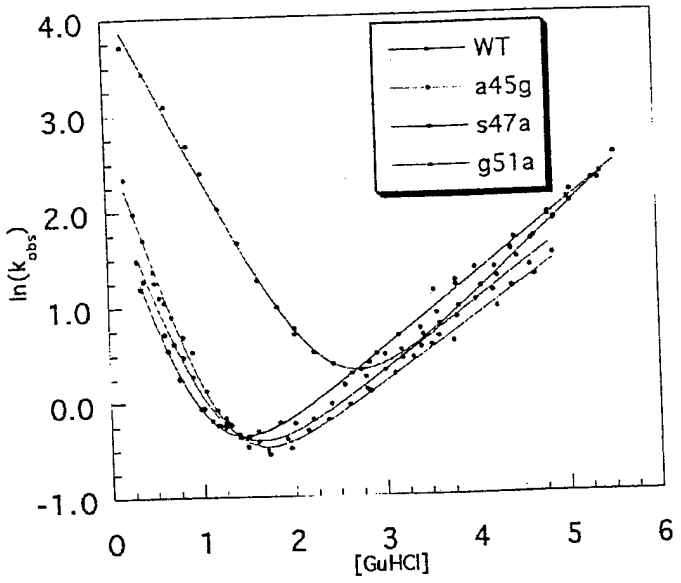
Distal loop hairpin. The distal loop hairpin consists of two β-strands connected by a tight β-turn (Fig.1*a*). Several mutations probe the integrity of the hairpin and reveal that it is relatively well structured in the transition state. Gly 51, located in the β-turn, has a positive φ angle which is disfavored for all amino acids except glycine; therefore, a substitution with alanine is likely to disrupt formation of the turn. The G51A mutation slowed the folding rate suggesting that the two strands joined by this turn are brought together at the transition state. Ser 47 and Thr 50 also contribute to the structure of the turn by forming local hydrogen bonds between their side chain hydroxyl oxygens and adjacent backbone amide protons (Fig.1*b*). Mutation of these residues significantly slowed the folding rate confirming the near-native structure of this part of the molecule in the transition state. Leu 44 and Tyr 55 interact with each other near the base of the hairpin on the more solvent exposed side; alanine substitutions at these positions slowed the folding rate and increased the unfolding rate, suggesting that the base of the hairpin is partially structured in the transition state.

Diverging type II β-turn. The diverging type II β-turn joining the RT and the n-src loops is stabilized by a hydrophobic interaction between two residues flanking the turn and a local hydrogen bond between the carboxylate of Glu 30 and a backbone amide proton (Fig.1). A local structure prediction program based on a library of recurrent sequence-structure motifs identified this region as the most likely portion of the protein to adopt structure in isolation, and a seven residue peptide corresponding to this turn has been found by NMR to be partially structured in solution[9]. Mutation of Glu 30 and Gly 29 to alanine affected both the folding and the unfolding rate suggesting that these residues have not fully formed all of their contacts in the transition state. However, the interpretation is somewhat complicated by the fact that these mutations are likely to disrupt residual structure in the denatured state.

Hydrophobic core. The hydrophobic residues we have characterized fall roughly into two classes. The first class consists of residues in a hydrophobic cluster formed by the base of the distal loop hairpin and the strand following the diverging turn. Mutations in all of these residues (Ala 45, Ile 34, Leu 32, Ile 56) significantly slowed the folding rate. The second class consists of residues outside of this hydrophobic cluster that are, for the most, partially solvent exposed. Mutations in these residues (Phe 10, Leu 24, Tyr 16, Val 61) had relatively small affects on the folding rate. Taken together, these results suggest that the hydrophobic interactions between the base of the distal hairpin and the strand following the diverging turn are at least partially formed in the folding transition state. The I34A mutation slows both the folding and the unfolding rate suggesting that it destabilizes the transition state more than the native state; the loss of interactions may be partially compensated by structural rearrangements or a relief of strain in the native state.

Hydrogen-bond network. Mutations in the hydrogen bond network residues (S47A, T50A and E30A) significantly reduce the folding rate. To our knowledge, this is the first example of the formation of a hydrogen bond cluster in a folding transition state. As all three residues are involved in both local and non-local hydrogen bonds in the native state it is hard to distinguish conclusively which interactions are important for stabilizing the transition state. However, the large effect of truncations of hydrophobic residues which pack between the distal loop hairpin and the diverging turn (L32A and A45G) on k_f suggests that the distal loop hairpin and diverging turn are not only well structured, but also closely opposed at the transition state. Therefore, it is likely that Glu 30 and Ser 47 are positioned in the proper geometry for the formation of a tertiary hydrogen bond.

Unstructured regions. The remainder of the src SH3 domain appears to be disordered in the transition state. Mutations in the RT loop (D15A, S18A), the n-src loop (N36A, G40A) and the cluster of surface aromatics either had no effect on kinetics



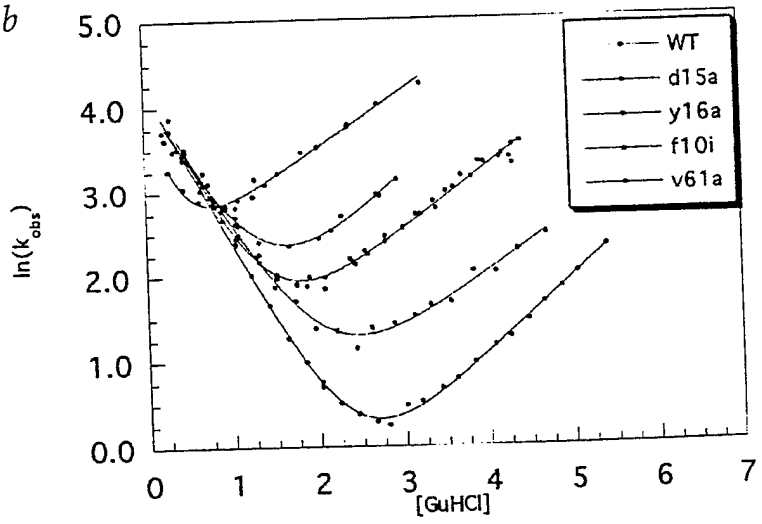


Fig. 2 Dependence of the rate of folding and unfolding on the denaturant concentration for ***a***, mutants which lower k_f and ***b***, mutants which increase k_u. The data for the wild type protein (black) is shown in both panels for comparison. For several of the mutants, the dependence of the folding rate on the guanidine concentration is greater than that of the wild type protein; these mutations may cause some expansion of the denatured state. The color scheme for the mutants is as follows: (*a*) A45G, red; S47A, blue; G51A, green: (*b*) Y16A, blue; F10I, green; V61A, magenta; D15A, red. The solid lines represent the fits to the experimental data.

or exclusively increased the unfolding rate, indicating that these regions do not play a role in guiding the protein towards its folded conformation. These structural elements are located on the opposite side of the molecule from the distal loop hairpin and the diverging turn and constitute the peptide binding site.

Structure of the folding SH3 transition state

To estimate the extent to which an interaction is formed in the transition state, it is convenient to normalize the effect of a mutation on the folding rate by its effect on overall stability (all other things being equal, it is expected that more drastic mutations will have larger effects). The ratio $(\Delta\Delta G_{U-\ddagger}/\Delta\Delta G_{U-F})$, termed the folding Φ_F value by Fersht and co-workers[8], conveniently describes the degree of structure formation around each residue in the transition state. Fig. 3*a* shows a schematic of the src SH3 structure[2] color-coded by Φ_F values on a continuous scale from 1 (red) to 0 (blue). Mutations with Φ_F values close to one (red) affect largely the folding, but not the unfolding rate, and are therefore likely to disrupt interactions stabilizing the transition state. Residues in the distal loop hairpin, diverging turn, and in the hydrophobic core between them have the highest Φ_F values and thus most native-like interactions in the transition state. Moving away from this high Φ_F (red) zone, the Φ_F values gradually decrease to 0 (blue) indicating that the remainder of the protein is largely unstructured in the transition state. The NMR study mentioned earlier showed that the diverging turn is at least partially formed in the denatured protein. The kinetic data reported here suggest that the rate-limiting step in folding involves formation of the distal loop β-hairpin and the docking of the hairpin onto the diverging turn and the strand following it. Once these elements are brought together the three-stranded β-sheet they form could serve as a nucleus around which the RT loop and the two terminal β-strands rapidly assume their native conformation.

NMR studies of the diverging β-turn peptide[9] and of the denatured state of the drk SH3 domain[10] suggest that both turns are partially sampled in the denatured state. If the rate limiting step in folding is a productive collision between these two structural elements, the length of the n-src loop, which connects the diverging turn and the distal loop, should impact the folding rate. The SH3 domain of PI3 kinase has an elongated n-src loop and folds significantly more slowly[11] than any of the other SH3 domains and several of the variants obtained in the SH3 simplification experiment[1] had deletions in the n-src loop. The strong complementarity of these structural elements is vividly illustrated by the recently published structure of the Eps8 SH3 domain dimer[12] in which the distal loop of one molecule is paired with the diverging turn of another.

The segments of the src SH3 domain found to be most structured in the transition state also include the residues most protected from hydrogen-deuterium (HD) exchange with the solvent[5]. There is not a residue-by-residue correspondence — mutagenesis probes side chain interactions, while HD exchange reports on hydrogen bonding and solvent accessibility of backbone amide protons. However, the distal hairpin and the diverging turn were indeed among the most highly protected parts of the protein. If the large fluctuations which lead to the unfolding transition state are amplified versions of the local fluctuations which contribute to HD exchange, the highest rates of exchange would be expected in parts of the protein most disrupted in the transition state[13]. A similar correlation has been observed for protein L[14], but not for CI-2[15].

Sequence conservation and Φ_F values

One of the goals of this study was to understand the sequence

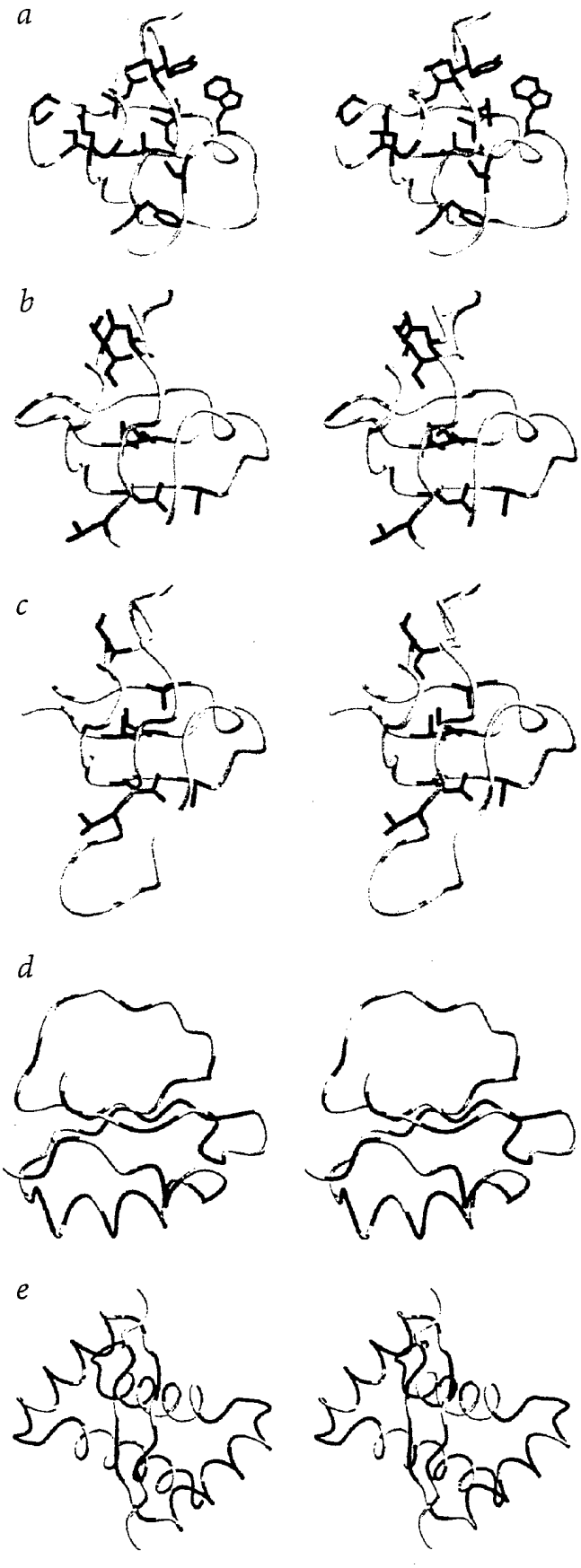


Fig. 3 ***a***, Structures of the src SH3 domain, ***b***, the wt α-spectrin SH3 domain and ***c***, the distal loop permutant, ***d***, CI2 and ***e***, the Arc repressor dimer colored by Φ_F value on a continuous scale from red (1) to blue (0). Experimentally determined values for CI2 were obtained from the original literature[22] and include only alanine substitutions. Φ_F values for Arc repressor were calculated using data in ref. 23 and 25. Two residues that probably make extensive interactions in the src SH3 folding transition state are not colored in (*a*) because of complications in data interpretation: Ile 34 appears to make stronger interactions in the transition state than in the native state (both the folding and unfolding rates are slowed), and Leu 32, which has a substantially decreased folding rate, is so destablized that an accurate Φ_F value could not be obtained. These residues are in the strand following the diverging turn. All images were created with MidasPlus[38,39].

constraints observed in previous phage selection experiments[1]. Two thirds of the conserved residues exhibited high Φ_F values. One exception was Leu 24 ($\Phi_F = 0.2$), however, mutation of this residue to alanine was found to substantially reduce the protein's affinity for the peptide substrate used in the selection (data not shown). When many residues are changed simultaneously, those important for kinetics could be conserved because they play important roles in specifying protein structure. A similar conclusion was reached in a lattice simulation study[16].

It was suggested previously, based on a comparison of Φ_F values for CI-2 and phylogenetic sequence variation within the family of small protease inhibitors homologous to CI-2[16], that residues involved in the rate-limiting step of folding are conserved in evolution. However, for proteins such as CI-2 in which hydrophobic core residues have the highest Φ_F values, it is difficult to disentangle selection for stability from selection for kinetics as core residues generally are the most critical for stability. The SH3 domain is particularly well suited for addressing the relationship between evolutionary conservation and Φ_F values because several residues which make important interactions in the transition state lie outside of the hydrophobic core. We find no relationship between phylogenetic variation and Φ_F values for the SH3 mutants described in this paper (data not shown). In fact, the opposite was observed both with and without exclusion of the binding residues: residues with high Φ_F values were somewhat less conserved than residues with low Φ_F values. The considerably greater opportunity for compensatory mutations in evolution (the replacement of the hydrogen bonding Glu 30, Ser 47 pair in the src SH3 domain by two hydrophobic residues in the PLC γ SH3 domain, for example) compared to the phage selection may account for the differences in the correlation between Φ_F values and sequence conservation in evolution and in the phage selection.

Topology and transition state structure

A previous study on the α-spectrin SH3 domain assessed the importance of topology in determining the structure of the folding transition state[17]. Circular permutants had very similar native structures, but displayed substantially different distributions of Φ_F values (Fig. 3*b*,*c*). The distal loop permutant was particularly affected, consistent with our results on the involvement of this structural element in the transition state. Changes in chain connectivity thus alter transition state structure. Our data on the src SH3 domain provides the first opportunity to examine the effect of sequence divergence on the detailed structure of the transition state (the src and α-spectrin SH3 domains are only 34% identical in sequence). The placement of the transition state along the reaction coordinate

Table 1 Positions conserved in the combinatorial mutagenesis selection

Position	% Burial	Amino acids[1] (% observed/% allowed)[2]				
Phe 10	76	Ile (0/70)	Val (67/10)	Leu (0/10)	Phe (33/10)	
Leu 24	69	Ile (0/70)	Val (0/10)	**Leu** (100/10)	Phe (0/10)	
Gly 29	22	Lys (0/25)	Glu (0/25)	Arg (0/25)	**Gly** (100/25)	
Glu 30	53	Lys (0/50)	**Glu** (100/50)			
Ala 45	99	Ile (0/25)	Val (6/25)	**Ala** (94/25)	Thr (0/25)	
Ser 47	79	Ile (0/25)	Val (0/25)	Ala (22/25)	Thr (6/25)	Asn (16/0) Ser (56/0)
Thr 50	21	Ala (0/25)	Gly (0/25)	Ser (50/25)	Thr (50/25)	
Gly 51	54	Lys (0/25)	Glu (0/25)	Arg (0/25)	**Gly** (100/25)	
Ile 56	99	Ile (67/70)	Val (33/10)	Leu (0/10)	Phe (0/10)	

[1]The residues conserved more than 90% in the selected variants are indicated in bold.
[2]The first number in the parentheses is the percent recovery of the residue in the phage selection experiments, the second number is the percent recovery expected in the absence of selective pressure given the design of the library.

is very similar for the fyn[18], src[5], and spectrin[19] SH3 domains (the m_f/m ratios for the three proteins are within experimental error), suggesting that their transition states may have similar structures. Comparison of our results with the effects of mutations in the α-spectrin SH3 domain[17,20] suggests that this is indeed the case. In earlier studies of point mutants in the spectrin SH3 domain, two mutants which reported on the proximity of the distal loop to the rest of the protein had the highest Φ_F values, and in the accompanying paper by Serrano and coworkers[21], an Asp to Gly substitution in the distal loop β-turn is shown to have a Φ_F value of 1. The similarity between the src and α-spectrin SH3 domain folding transition states (compare Fig. 3a and 3*b*) is remarkable given their significant divergence in sequence and suggests that the exact identities of the amino acids are not important for the structure of the transition state. In fact, the critical hydrogen bonding residue Ser 47 in the distal loop of src SH3 is replaced in α-spectrin by a valine which contributes to the hydrophobic core. Taken together, the observations that the structure of the transition state is (i) altered by changes in topology (circular permutants), but (ii) largely invariant to the large number of substitutions between the SH3 domains strongly support the idea that topology is a dominant determinant of the folding mechanism of this family of proteins.

Previous work on CI-2[22] and the P22 Arc repressor[23] suggested that the transition states of these proteins represent expanded forms of the native state with most interactions partially formed. The picture for src SH3 domain is quite different: its transition state appears to be quite polarized, with one portion of the molecule much more highly ordered than the rest. For both CI-2 and the Arc repressor, the extent to which a residue contributes to the stability of the transition state is roughly proportional to its contribution to native state stability, as evidenced by the linear relationship between $\Delta\Delta G_U$ and $\Delta\Delta G_{U-\ddagger}$ observed in Brønsted plots[22,23]. In contrast, such a plot for the src SH3 mutants (data not shown) is more similar to that of barnase[22,24], a larger protein that folds through an intermediate: the data are scattered between lines with slopes of 0 and 1 indicating non-uniform formation of structure in the transition state.

Comparison of the structures of CI-2 and Arc repressor colored by Φ_F value (Fig. 3*d*,*e* respectively) with that of the src SH3 domain (Fig. 3*a*) accentuates the difference between the transitions states of these proteins. Both CI-2 and Arc repressor show a relatively uniform distribution of low (blue) and intermediate (magenta) Φ_F values, while the src SH3 domain is split into a high Φ_F value region (red) and a low Φ_F value (blue) one. Although a greater number of mutants was generated for CI-2 and the Arc repressor, the absolute number of residues with Φ_F values greater than 0.5 is still larger for src SH3 than for the two other proteins: six for src SH3 domain *versus* one for Arc repressor[23,25] and two for CI2[22]. The unusual features of the src SH3 transition state may reflect the dominant effect of topology in specifying the transition state structure. Unlike the other proteins whose transition states have been characterized, the SH3 domain consists predominantly of β-sheets. The rate-limiting step in folding may involve docking of transiently formed structural elements; in β-sheet proteins such local structure is likely to occur in the vicinity of β-turns, and in helical proteins, near the middle of helices. Thus, the folding transition states of β-sheet proteins may be expected to be more polarized and not as centered on the hydrophobic core as those of helix containing proteins. More generally, the topology of β-sheet proteins is to some extent determined by the positions and changes in chain orientation in the β-turns (in the SH3 domain, the diverging turn is one of the two transitions between the sheets, and the distal loop β-turn sets up the hydrophobic contacts along the distal loop β-hairpin), and thus formation of a small number of critical β-turns could be coupled to the formation of sufficient favorable native interactions to overcome the entropic barrier to folding. The polarization may also reflect the importance of hydrogen bonds in stabilizing the transition state: hydrogen bonds have much stronger orientational constraints than hydrophobic interactions and are not likely to be stabilizing unless almost fully formed.

In the past several years there has been considerable discussion of the differences between the 'new' and the 'classical' views of protein folding[26–29]. For small, single domain proteins with kinetics and thermodynamics well described by a two-state model, the distinction primarily concerns the breadth of the transition state ensemble: in the classical view, the transition state consists of a relatively well defined set of conformations, whereas for the funnel shaped energy landscapes suggested by the new view, the set of conformations can be extremely diverse (in the limit of the models of ref. 6, the transition state ensemble includes all conformations with a particular degree of order). It is important to note that the transition state approximation is valid independent of the homogeneity (or lack thereof) of the transition state, and that both old and new views are consistent with the simple exponential kinetics observed for the folding of small proteins. In the sequence simplification experiments[1], the folding rate was relatively

Table 2 Kinetic and thermodynamic parameters for wt SH3 domain and mutants[1]

Name	ΔG^{H_2O} (kcal mol^{-1})	$k_f^{0.3}$ (s^{-1})	$k_u^{3.5}$ (s^{-1})	$\Phi_F^{H_2O}{}_{kin}$	$\Phi_F^{H_2O}{}_{eq}$	$\Phi_F^{0.3}{}_{Cm}$
WT	3.7	39	2.1	—	—	—
F10I	2.7	39	32	0.05	0.08	0.00
D15A	3.3	42	5.5	0.01	0.02	-0.09
Y16A	0.9	24	91	0.10	0.11	0.08
S18A	4.1	46	1.2	*[2]	*	*
L24A	1.9	17	21	0.21	0.25	0.20
G29A	2.1	12	10	0.33	0.39	0.29
E30A	1.8	4.6	7.4	0.79	0.67	0.67
L32A	0.1	4.8	9.6	^[3]	^	^
I34A	3.0	5.1	0.4	5.5	1.61	1.31
N36A	4.0	45	3.3	*	*	*
G40A	3.8	20	1.8	*	*	*
W42A	2.5	27	12	0.07	0.08	0.11
L44A	1.5	7.8	9.3	0.37	0.35	0.34
A45G	2.2	6.9	1.8	0.72	0.58	0.54
S47A	1.2	2.9	2.4	0.87	0.60	0.75
T50A	1.9	3.2	3.5	0.73	0.72	0.57
G51A	2.0	4.7	1.9	0.77	0.67	0.64
Y55A	1.9	8.8	7.5	0.39	0.43	0.35
I56A	1.4	3.9	4.7	0.58	0.49	0.49
Y60A	4.1	40	1.2	*	*	*
V61A	3.1	40	18	-0.01	-0.01	-0.01

[1]All experiments were performed at 295 K using 50 mM sodium phosphate as buffer. $k_f^{0.3}$ is the folding rate in 0.3 M guanidine, $k_u^{3.5}$, the unfolding rate in 3.5 M guanidine. Details on the three different methods for calculating Φ_F values are described in the Methods section.
[2]These mutants (*) were more stable than wt SH3.
[3]Due to the very low stability of this mutant (^) thermodynamic and kinetic parameters are only rough estimates.

unchanged by drastic changes in the sequence, consistent with a simple funnel picture in which interactions stabilizing the native state also stabilize partially folded conformations. The clustering of mutations which primarily affect the folding rate (have high Φ_F values) in the distal loop and diverging turn suggest that the transition state ensemble for the src SH3 domain is relatively well defined and thus that the folding funnel departs considerably from symmetry. Taken together, our data suggest that the folding free energy landscape of the src SH3 domain is somewhere between that envisioned in the classical view of folding and the extreme of a completely symmetrical funnel.

Protein folding landscapes could in principle deviate from symmetry either because of heterogeneities in inter-residue contact energies or because of asymmetries in the folded structure. The robustness of the SH3 domain transition state to large changes in sequence (and hence changes in the residue-residue interaction energies) indicated by the strong similarities between the src and spectrin SH3 transition states and the near wild type folding rates of the simplified SH3 variants, and the contrast with the more delocalized transition states of Arc repressor and CI-2 suggest that topological features of the SH3 domain fold rather than heterogeneities in the contact energies are likely to be responsible for the departure from symmetry. The importance of topology in determining folding mechanism is further highlighted by a comparison to λ repressor[30]; two relatively subtle Gly to Ala substitutions in one of the helices of this protein caused a much larger change in m_f/m (from 0.4 to 0.8) than the very large number of sequence changes between src SH3, fyn SH3, spectrin SH3, and the simplified SH3 variant FP1 (ref. 1) (m_f/m = 0.69, 0.68, 0.69 and 0.63, respectively). By analogy with these results, differences in topology may also underlie the differences in folding scenarios derived from studies of lattice models of proteins (the delocalized nuclei of ref. 31 versus the specific nucleus of ref. 32, for example; it was anticipated that the delocalized nuclei scenario may be more common for small helical proteins[31]). The recent finding that the folding rates and m_f/m ratios for small single domain proteins are correlated strongly with the average separation between contacting residues suggests that the relationship between topology and folding mechanism is quite general[33].

Our data present a challenge for methods that seek to predict folding transition state structure using molecular dynamics[34] and other computational approaches[31,35]. Agreement between theory and experiment has been claimed in a number of cases involving CI2, but for this protein it is only necessary to predict that most interactions are partially formed to achieve reasonable success. The highly polarized src SH3 transition state will provide a much more rigorous test of computational models as it requires the precise identification of crucial residues. To make at least part of the test blind, we are currently determining Φ_F values for the remaining residues in the structure and invite predictions of these as a test of computational models of protein folding.

Methods

Mutagenesis and purification. The SH3 gene was cloned into the NdeI and BamHI sites of the pET 15b expression vector (Novagen). Mutagenesis was accomplished using the Quick Change Site-Directed mutagenesis kit (Stratagene). Plasmids harboring the point mutations were transformed into BL21 cells, and protein was overexpressed and purified[5]. The His•Tag® was not removed for the purposes of this study. All mutants were sequenced to ensure that the mutagenesis was successful and the purified proteins were analyzed by mass spectrometry to confirm that each mutation was the expected one.

Biophysical analysis. In all experiments, proteins solutions were made in 50 mM sodium phosphate (JT Baker), pH 6, and the temperature was held constant at 295 K. The stability of the point mutants was assessed by guanidine denaturation using either CD or fluorescence as described[36]. The kinetics of folding and unfolding were followed by fluorescence on a Bio-Logic SFM-4 stopped-flow instrument[36]. The unfolding reaction for the wild type protein was well modeled as a two-state process [5], and the kinetic and equilibrium data for the mutants were fit to a two-state model.

Φ value analysis. There are several different ways of measuring the values of $\Delta\Delta G_{U\text{-}F}$ and $\Delta\Delta G_{U\text{-}‡}$ that determine Φ_F. Because of the possible errors introduced by extrapolation we report three estimates of the Φ_F value for all the mutants destabilized by more than 0.5 kcal mol^{-1}: (i) for $\Phi_F^{H_2O}{}_{kin}$, both $\Delta\Delta G_{U\text{-}F}$ and $\Delta\Delta G_{U\text{-}‡}$ were computed from kinetic data extrapolated to H_2O[15]; (ii) for $\Phi_F^{H_2O}{}_{eq}$, $\Delta\Delta G_{U\text{-}F}$ was computed from equilibrium data and $\Delta\Delta G_{U\text{-}‡}$ from kinetic data extrapolated to H_2O; (iii) $\Phi_F^{0.3}{}_{Cm}$, $\Delta\Delta G_{U\text{-}F}$ was computed using (ΔC_m m_{avg}) and $\Delta\Delta G_{U\text{-}‡}$ from the folding rate in 0.3 M

guanidine[37]. The values obtained by the three methods match very closely confirming the validity of our results. Small differences are seen only in the significantly destabilized mutants for which estimates of the equilibrium ΔG_U are not very accurate due to the lack of a folded baseline.

Acknowledgments

We thank M. Eck for providing us with the atomic coordinates of src tyrosine kinase prior to submitting them in the Brookhaven protein data bank, Q. Yi for mass spectrometry analysis of all the SH3 mutants, J. Onuchic and members of the Baker group for useful comments on the manuscript, and L. Serrano and coworkers for sharing their manuscript on the spectrin folding transition state prior to publication. This work was supported by a grant from the Office of Naval Research and Young Investigator awards to D. B. from the NSF and the Packard Foundation.

Received 11 March, 1998; accepted 25 June, 1998.

1 Riddle, D. S., Santiago, J. V., Bray, S. T., Doshi, N., Grantcharova, V. P. & Baker, D. Functional rapidly folding proteins from simplified amino acid sequences. *Nature Struct. Biol.* **4**, 805–809 (1997).
2. Xu, W., Harrison, S. C., & Eck, M. J. Three-dimensional structure of the tyrosine kinase c-Src. *Nature* **385**, 595–602 (1997).
3. Yu, H., Rosen, M. K., & Schreiber, S. L. ^{1}H and ^{15}N assignments and secondary structure of the Src SH3 domain. *FEBS Lett.* **324**, 87–92 (1993).
4. Pace, N. C., Shirley, B. A., Mcnutt, M. & Gajiwala, K. Forces contributing to the stability of proteins. *FASEB* **10**, 75–83 (1996).
5. Grantcharova, V. P. & Baker, D. Folding dynamics of the src SH3 domain. *Biochemistry* **36**, 15685–15692 (1998).
6. Doyle, R., Simons, K., Qian, H. & Baker, D. Local interactions and the optimization of protein folding. *Proteins Struct. Func. Gen.* **29**, 282–291 (1997).
7. Socci, N. D., Onuchic, J. N. & Wolynes, P. G. Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* **104**, 5860–5868 (1996).
8. Fersht, A. R. Characterizing transition states in protein folding: an essential step in the puzzle. *Curr. Opin. Struct. Biol.* **5**, 79–84 (1994).
9. Yi, Q., Bystroff, C. & Baker, D. Prediction and structure characterization of an independently folding substructure in the src SH3 domain. *J. Mol. Biol.*, **in the press** (1998).
10. Zhang, O. & Forman-Kay, J. D. NMR studies of unfolded states of an SH3 domain in acqueos solution and denaturing conditions. *Biochemsitry* **36**, 3959–3970 (1997).
11. Kishan, K. V. R., Scita, G., Wong, W. T., Di Fiore, P. P. & Newcomer, M. E. The SH3 domain of Eps8 exists as a novel intertwined dimer. *Nature Struct. Biol.* **4**, 739–743 (1997).
12. Guijarro, J. I., Morton, C., Plaxco, K. W., Campbell, I. D. & Dobson, C. M. Folding kinetics of the SH3 domain of PI3 kinase by real-time NMR combined with optical spectroscopy. *J. Mol. Biol.* **276**, 657–667 (1998).
13. Woodward, C. Is the slow-exchanging core the protein folding core? *TIBS* 18, 359–**360** (1993).
14. Gu, H., Kim, D. & Baker, D. Contrasting roles for the symmetrically disposed β-turns in the folding of a small protein. *J. Mol. Biol.* **274**, 588–596 (1997).
15. Neira, J. L., Itzhaki, L. S., Otzen, D. E., Davis, B. & Fersht, A. R. Hydrogen exchange in chymotrypsin inhibitor 2 probed by mutagenesis. *J. Mol. Biol.* **270**, 1–12 (1997).
16. Shakhnovich, E., Abkevich, V. & Ptitsyn, O. Conserved residues and the mechanism of protein folding. *Nature* **379**, 96–98 (1996).
17. Viguera, A.R., Serrano, L. & Wilmanns, M. Different folding transition states may result in the same native structure. *Nature Struct. Biol.* **3**, 874–879 (1996).
18. Plaxco, K.W., Guijarro, J. I., Morton, C. J., Pitkeathly, M., Campbell, I. D. & Dobson, C. M. The folding kinetics and thermodynamics if the fyn SH3 domain. *Biochemistry* **37**, 2529–2537 (1998).
19. Viguera, A. R., Martinez, J. C., Filimonov, V. V., Mateo, P. L., & Serrano, L. Thermodynamic and kinetic analysis of the SH3 domain of spectrin shows a two-state folding transition. *Biochemistry* **33**, 2142–2150 (1994).
20. Prieto, J., Wilmans, M., Jimenez, M. A., Rico, M. & Serrano, L. Non-native interactions in protein folding and stability: introducing a helical tendency in the all β-sheet α-spectin SH3 domain. *J. Mol. Biol.* **268**, 760–778 (1997).
21. Martinez, J. C., Pisabarro, M. T. & Serrano, L. Obligatory steps in protein folding and conformational diversity of the transition state. *Nature Struct. Biol.* **5**, 721–729 (1998).
22. Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. The structure of the transition state for folding of chymotrypsin inhibitor 2 analyzed by protein engineering methods: evidence for a nucleation condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260–288 (1995).
23. Milla, M. E., Brown, B. M., Waldburger, C. D. & Sauer, R. T. P22 Arc Repressor: Transition state properties inferred from mutational effects on the rates of protein unfolding and refolding. *Biochemistry* **34**, 13914–13919 (1995).
24. Serrano, L., Matouschek, A. & Fersht, A. The folding of an enzyme. 3. Structure of the transition state for unfolding of barnase analyzed by a protein engineering procedure. *J. Mol. Biol.* **224**, 805–818 (1992).
25. Milla, M. E., Brown, B. M. & Sauer, R. T. Protein stability effects of a complete set of alanine substitutions in Arc repressor. *Nature Struct. Biol.* **1**, 518–523 (1994).
26. Baldwin, R. L. Matching speed and stability. *Nature* **369**, 183–184 (1994).
27. Wolynes, P. G., Onuchic, J. N. & Thirumalai, D. Navigating the folding routes. *Science* **267**, 1619–20 (1995).
28. Dill, K. A. & Chan, H. S. From Levinthal to pathways to funnels. *Nature Struct. Biol.* **4**, 10–19 (1997).
29. Pande, V.S., Grosberg, A.Y., Tanaka, T. & Rokhsar, D. S. Pathways for protein folding:is a new view needed? *Curr. Opin. Struct. Biol.* **8**, 68–79 (1998).
30. Burton, R. E., Huang, G. S., Daugherty, M. A., Calderone, T. L. & Oas, T. G. the energy landscape of a fast-folding protein mapped by Ala→Gly substitutions. *Nature Struct. Biol.* **4**, 305–310 (1997).
31. Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z. & Wolynes, P. G. Protein folding funnels: the nature of the transition state ensemble *Folding Design* **1**, 441–450 (1996).
32. Abkevich, V. I., Gutin, A.M. & Shakhnovich, E. I. Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33**, 10026–10036 (1994).
33. Plaxco, K. W., Simons, K. & Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994 (1998).
34. Li, A. & Daggett, V. Identification and characterization of the unfolding transition state of chymotypsin inhibitor 2 by molecular dynamics simulations. *J. Mol. Biol.* **257**, 412–429 (1996).
35. Schoemaker, B. A., Wang, J. & Wolynes, P. G. Structural correlations in protein folding *Proc. Natl. Acad. Sci. USA* **95**, 777–782 (1997).
36. Scalley, M. L. *et al.* Kinetics of folding of the IgG binding domain of peptostreptoccocal protein L. *Biochemistry* **36**, 3373–3382 (1996).
37. Jackson, S. E., Moracci, M., elMarsy, N., Johnson, C. & Fersht, A.R. Effect of cavity-creating mutations in the hdrophobic core of chymotrypsin inhibitor 2. *Biochemistry* **32**, 11259–11269 (1993).
38. Ferrin, T. E., Huang, C. C., Jarvis, L. E. & Langridge, R. The MIDAS display system. *J. Mol. Graph.* **6**, 13–27 (1988).
39. Huang, C. C., Pettersen, E. F., Klein, T. E., Ferrin, T. E. & Langridge, R. Conic: a fast renderer for space-filling molecules with shadows. *J. Mol. Graph.* **9**, 230–236 (1991).

Article No. mb982072 J. Mol. Biol. (1998) 283, 293–300

JMB

Prediction and Structural Characterization of an Independently Folding Substructure in the src SH3 Domain

Qian Yi, Chris Bystroff, Ponni Rajagopal, Rachel E. Klevit and David Baker*

Department of Biochemistry University of Washington Seattle, WA 98195, USA

*Corresponding author

Previous studies of the conformations of peptides spanning the length of the α-spectrin SH3 domain suggested that SH3 domains lack independently folding substructures. Using a local structure prediction method based on the I-sites library of sequence-structure motifs, we identified a seven residue peptide in the src SH3 domain predicted to adopt a native-like structure, a type II β-turn bridging unpaired β-strands, that was not contained intact in any of the SH3 domain peptides studied earlier. NMR characterization confirmed that the isolated peptide, FKKGERL, adopts a structure similar to that adopted in the native protein: the NOE and $^3J_{NH\alpha}$ coupling constant patterns were indicative of a type II β-turn, and NOEs between the Phe and the Leu side-chains suggest that they are juxtaposed as in the prediction and the native structure. These results support the idea that high-confidence I-sites predictions identify protein segments that are likely to form native-like structures early in folding.

Keywords: SH3 domain; folding initiation site; protein folding; local structure; β-turn

Introduction

There has been considerable discussion about the role of local interactions in protein folding (Abkevich *et al.*, 1995; Avbelj & Moult, 1995; Doyle *et al.*, 1997; Fersht, 1995; Muñoz & Serrano, 1996; Unger & Moult, 1996). Direct experimental studies are hampered because early events of protein folding are likely to take place within the microsecond time-scale. To investigate the role of β-turns, peptide fragments derived from proteins have been studied as models for early events in protein folding (Dyson *et al.*, 1988, 1992). Those studies have shown that sequences that form turns in proteins can also form reasonably stable turns in short peptides in aqueous solution. However, because non-local interactions play an important role in stablizing structure in both the native state and denatured state (Wang & Shortle, 1997), many peptides derived from proteins do not adopt well-defined structure in isolation. In particular, a recent study of five peptides derived from the α-spectrin SH3 domain failed to detect any persistent structure and it was concluded that folding initiation sites do not play a role in the folding of this all-β protein (Viguera *et al.*, 1996).

We have recently developed a method for local protein structure prediction based on a library (I-sites) of 7 to 19 residue sequence patterns that strongly correlate with local protein structural features (Bystroff & Baker, 1997, 1998). The sequence segments matching a particular pattern almost always adopt the same conformation in a wide range of protein structures, suggesting that local interactions within the segment are strong enough to override the differences in non-local interactions. Inspection of the sequence-structure motifs in most cases readily reveals the interactions that stabilize the observed structure. One of the novel motifs is a "diverging type II β-turn" stabilized by a side-chain-to-backbone hydrogen bond and a pair of inwardly turned hydrophobic residues bracketing the turn (Bystroff & Baker, 1998). The turn is

Abbreviations used: 1D, one-dimensional; ppb, parts per billion; ppm, parts per million; NOE, nuclear Overhauser effect; TOCSY, total correlated spectroscopy; ROESY, rotating frame Overhauser effect spectroscopy; TSP, 3-(Trimethylesilyl)propionic-2,2,3,3-d_4 acid; MALDI-MS, matrix assisted laser desorption ionization mass spectrometry.

E-mail address of the corresponding author: baker@ben.bchem.washington.edu

referred to as diverging because the two strands connected by it do not form backbone hydrogen bonds.

We have proposed that, since the interactions within the sequence segments override non-local interactions, peptide segments that closely match one of the sequence patterns are likely to adopt structure in isolation and potentially serve as folding initiation sites. Because the SH3 domain had been proposed to lack such peptide segments, we were curious about I-sites predictions of local structure for this protein family. Interestingly, the peptide segment predicted to be the most likely to have structure in isolation was a diverging type II β-turn that was not contained intact in any of the α-spectrin SH3 peptides studied previously (Viguera *et al.*, 1996). In this study, we demonstrate that the sequence FKKGERL, derived from the diverging type II turn in the src SH3 domain, adopts that conformation in isolation.

Results

Local structure prediction

Figure 1(a) shows the location of all correct (red) and incorrect (blue) fragment structure predictions (see Materials and Methods) for the src SH3 domain. The two highest-confidence predictions occurred around the position of the type II β-turn at residues 27 to 30 (shaded region in Figure 1(b)); both correctly predict a type II turn, an inwardly turned glutamate side-chain, and the juxtaposition of the Phe26 and Leu32 side-chains. We selected the peptide FKKGERL (residues 26 to 32) for structural studies, omitting three residues of the β-strand. The high-confidence prediction is a consequence of the similarity between the sequence pattern for the I-sites diverging turn motif (Figure 2(a)) and the SH3 multiple sequence alignment at residues 26 to 32 (Figure 2(c)). As expected, given the strong similarity of the sequence patterns, the predicted structure (Figure 2(b)) is very similar to that in the crystal structure of the src SH3 domain (Figure 2(d)).

NMR study of the peptide conformation

To avoid artifacts due to non-native electrostatic interactions between the N and C termini, we studied the structure of two forms of the peptide; one with free N and C termini, the other with an acetylated N terminus and an amidated C terminus. The results were virtually identical for the two peptides; for brevity, we present only the data for the blocked peptide.

Backbone conformation

The proton NMR spectrum of the peptide was completely assigned using a 50 ms TOCSY experiment and confirmed by a 250 ms ROESY experiment. The chemical shifts for all proton resonances

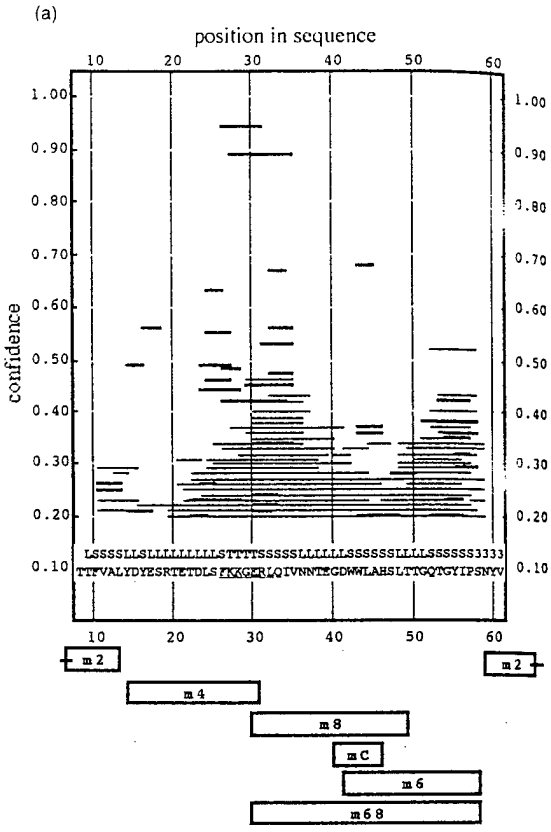


Figure 1. (a) I-sites fragment predictions for the SH3 domain sequence family. Red bars indicate correct predictions according to the crystal structure (1FMK); blue bars are incorrect predictions. In the text box is the sequence and secondary structure of the src SH3 domain; (S, strand; T, turn; L, loop; 3, 3_{10} helix). The peptides of α-spectrin SH3 domain studied by Viguera *et al.* (1996) are indicated by the rectangular bars below the graph. (b) Backbone trace of the SH3 domain (1FMK) showing the location of the diverging turn (shaded).

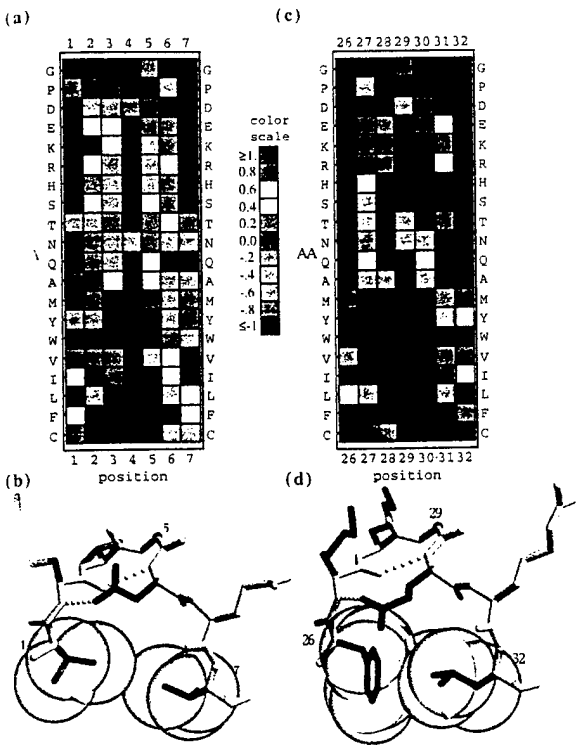


Figure 2. (a) Sequence profile and (b) paradigm structur (EFT residues 247 to 253) for the diverging turn motif from the I-sites library (Bystroff & Baker, 1998). (c) Sequence profile for the diverging turn segment in the SH3 domain sequence family, and (d) the native structure of this portion of the src SH3 domain (1FMK residues 102 to 108; Xu *et al.*, 1997). Numbering in (c) and (d) is consistent with our previous studies (Riddle *et al.*, 1997). The colors in the profile tables represent the log likelihood ratio, log $[P_{ij}/P_i]$, where P_{ij} is the frequency of amino acid i at position j in the I-sites sequence pattern (a) or SH3 domain multiple sequence alignment (c), and P_i is the average frequency of amino acid i in the proteins. The color scale is shown in the central panel; favored amino acids are in red and disfavored in blue. Conserved polar residues in the sequence profiles are shown in green; conserved non-polar residues in purple. Solvent-accessible surfaces are drawn around the non-polar side-chains. The Figure was made using Raster3D (Merritt & Murphy, 1994) and Mathematica (Wolfram Research, Inc).

are summarized in Table 1. The NOE pattern observed in ROESY spectra, the three-bond NH^{α} coupling constants, and the temperature coefficients of the amide protons are summarized in Figure 3. As described in the following paragraphs, all the NMR parameters indicate that a type II β-turn conformation is significantly populated.

β-Turns yield characteristic patterns of sequential and medium-range NOEs and $^3J_{NH_\alpha}$ coupling constants (Wüthrich, 1986). NOE patterns indicative of β-turns include a d_{NN} NOE between residues 3 and 4 and a $d_{\alpha N}(i, i+2)$ NOE between residues 2 and 4 in the turn. Type I and II turns are distinguished by the relative strengths of the $d_{\alpha N}(2,3)$ NOE (stronger for type II) and the $d_{NN}(2,3)$ NOE (stronger for type I). A strong $d_{NN}(i, i+1)$ NOE between Gly29 and Glu30, and a less strong $d_{\alpha N}(i, i+2)$ NOE between Lys28 and Glu30 in the peptide were observed in ROESY spectra (Figure 4(a)). The presence of this characteristic NOE pattern suggests a high preference for a β-turn conformation for the Lys-Lys-Gly-Glu segment in the peptide. The pattern of a strong $d_{\alpha N}$ and a weak d_{NN} NOE between Lys28 and Gly29, along with a very weak $d_{\beta N}$ NOE between Lys28 and Glu30 (observed at lower thresholds, but not shown in Figure 4(a)), is most consistent with a type II β-turn (Campbell *et al.*, 1995; Wüthrich, 1986). β-Turns are characterized also by a small value (~5 Hz) of $^3J_{NH_\alpha}$ for residue 2 in the turn (Wüthrich, 1986). The observed $^3J_{NH_\alpha}$ coupling constant for Lys28 (5.0 Hz, Figure 3), compared to the $^3J_{NH_\alpha}$ value of 6.6 Hz for Lys in the random coil conformation (Smith *et al.*, 1996), is consistent with its position as residue 2 in a turn conformation.

In most β-turns there is a backbone-backbone hydrogen-bond between the carbonyl oxygen atom of residue 1 and the NH group of residue 4. This hydrogen bond often leads to a small temperature coefficient $(0 < -\Delta\delta/\Delta T < 5$ ppb $K^{-1})$ for the amide proton of residue 4 (Rose *et al.*, 1985) . The amide proton of Glu30 has a temperature coefficient of -4.6 ppb K^{-1} (Figure 3) (compared to $6 < -\Delta\delta/\Delta T < 10$ ppb K^{-1} for amide groups in a random coil conformation), suggesting a role as a hydrogen-bond donor, as expected for residue 4 in a turn conformation. This, combined with the evidence for a type II turn conformation, would implicate the carbonyl oxygen atom of Lys27 as the likely H-bond acceptor.

Table 1. Proton assignments in 50 mM sodium phosphate at pH 6.0 and 12°C

Residue	NH	$C^{\alpha}H$	$C^{\beta}H$	Others
Phe26	8.34	4.54	3.09, 3.01	δ7.26, ε7.36, ζ7.32
Lys27	8.45	4.30	1.78, 1.68	γ1.38, δ1.65, ε2.99
Lys28	8.48	4.20	1.80, 1.73	γ1.48; 1.45, δ1.71, ε3.03
Gly29	8.65	3.93, 3.99		
Glu30	8.04	4.31	1.95, 2.04	γ2.09; 2.25
Arg31	8.55	4.31	1.84, 1.78	γ1.61, δ3.19, $N^{\varepsilon}H$ 7.25
Leu32	8.41	4.31	1.67, 1.60	γ1.62, δ0.93; 0.87

TSP was used as a reference for chemical shifts.

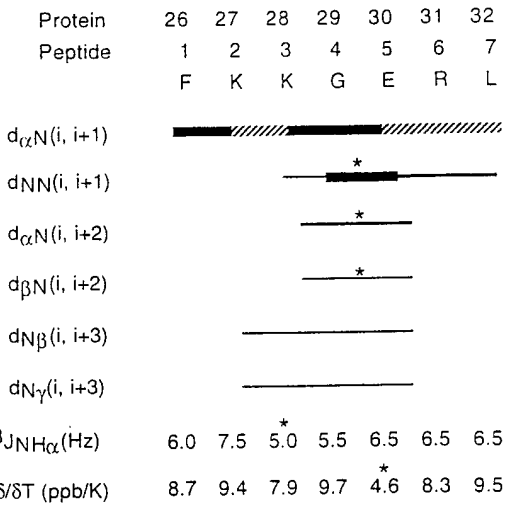


Figure 3. Observed NOE patterns, values of the $^3J_{NH\alpha}$ coupling constants, and the temperature coefficients for the peptide FKKGERL. The widths of the lines represent the NOE intensities. The broken lines represent some of the $d_{\gamma N}(i-1, i)$ NOEs that overlap with $d_{\gamma N}(i, i)$ NOEs. The results indictive of the type II β-turn conformation are indicated with asterisks (*).

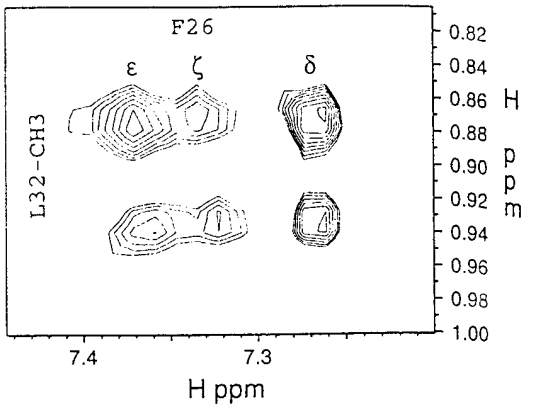


Figure 5. Portion of the ROESY spectrum displaying the NOEs between the aromatic protons of Phe26 and the methyl groups of Leu32 at 12 C and pH 6.0.

Side-chain interactions that stabilize the β-turn

In addition to the backbone hydrogen bond (Lys27 CO-Glu30 NH) interaction described above, there are two other interactions observed in the folded protein that may contribute to the stability of the isolated β-turn (Figure 2(d)). The first is a hydrophobic interaction between the aromatic ring of Phe26 and the methyl groups of Leu32. NOE cross-peaks between the aromatic protons of Phe26 and the methyl groups of Leu32 are clearly visible in the ROESY spectrum at 12 C (Figure 5), indicating that conformations are populated in which the aromatic ring and the methyl groups are in relatively close contact (<5 Å). The side-chain NOE between Phe26 and Leu32 is observed at 12 C but

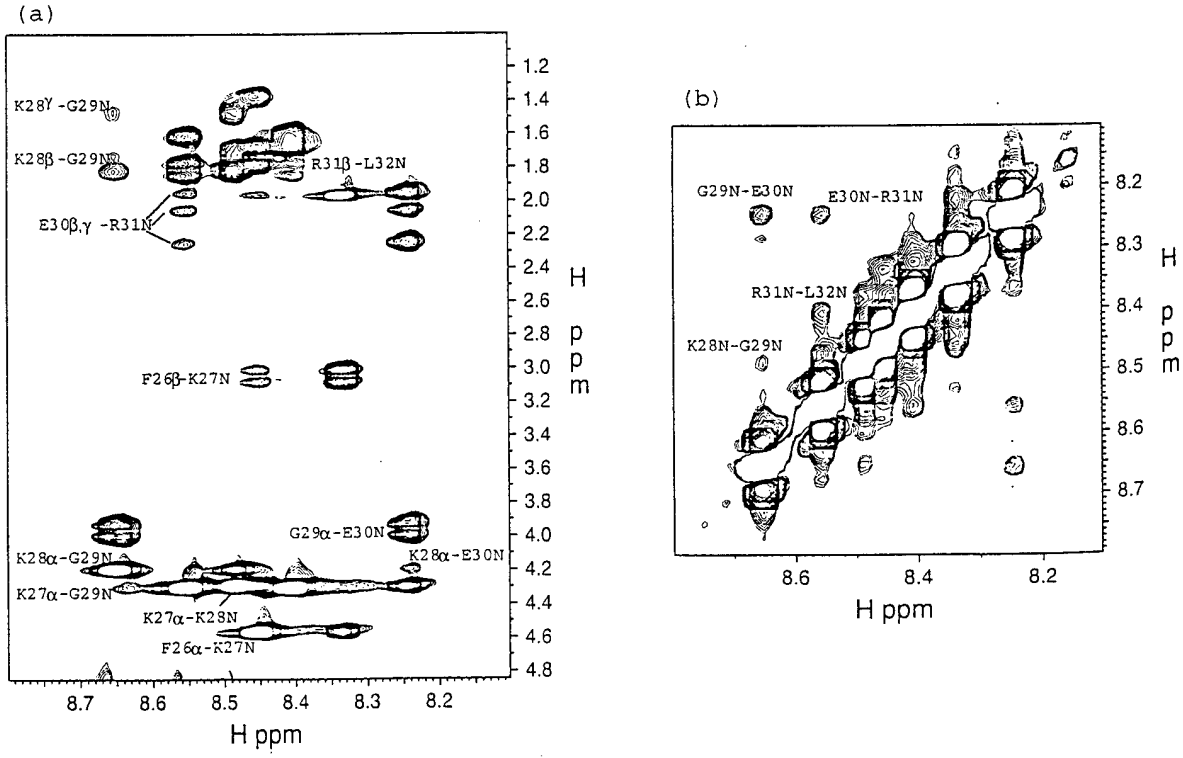


Figure 4. ROESY spectrum at pH 6.0 and 12 C. (a) the C(α, β)H - NH region, (b) the NH - NH region. A 250 ms mixing time was used. Only the inter-residue NOEs are labelled in the spectra.

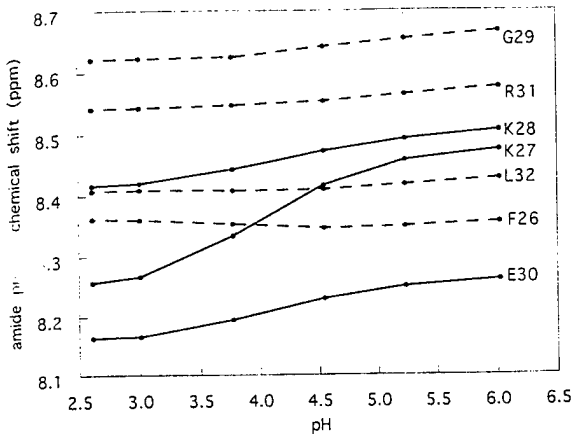


Figure 6. Dependence of the amide proton chemical shifts on pH. The amide protons of K27, K28 and E30 (continuous curves)are sensitive to pH, while the amide groups of F26, G29 and L32 (broken curves) are not.

not at 24°C (data not shown). These results suggest that the two hydrophobic side-chains are juxtaposed as in the prediction and in the native structure.

The second native interaction is a hydrogen bond between a side-chain carboxylate oxygen atom of Glu30 and the backbone amide proton of Lys27. We reasoned that such an interaction would be dependent on the ionization state of the Glu30 side-chain, the only ionizable group in the peptide with a pK_a between 2.5 and 6.5, and carried out a pH titration to investigate this possibility. The amide proton of Lys27 undergoes the largest chemical shift change of any amide over a pH range of 2.5 to 6.5. When the Glu30 side-chain is deprotonated at pH 6.0, the amide proton of Lys27 in the peptide is shifted downfield by ~0.22 ppm relative to pH 2.6 (Figure 6). This shift is consistent with an interaction with the Glu30 side-chain, as hydrogen-bonding will cause deshielding of the amide proton, which generally results in a downfield shift for the amide proton. The temperature coefficient of Lys27 amide proton is greater than the expected value for an amide proton involved in hydrogen-bonding in proteins (-9.4 ppb K^{-1} *versus* -5.0 ppb K^{-1}) but, since an increase in temperature is likely to increase the mobility of the Glu30 side-chain significantly, the temperature coefficient is probably not an accurate indicator of side-chain to main-chain hydrogen bond formation in the peptide. Weak NOEs between the β and γ protons of the Glu30 side-chain and the amide proton of Lys27 were observed at low threshholds in ROESY spectra (Figure 3). These results further suggest that the side-chain of the Glu30 is pointed toward the Lys27 amide proton, consistent with a hydrogen bond between the Lys27 amide proton and the side-chain carboxyl oxygen atom of Glu30.

Discussion

The structure of fragments of the src-SH3 domain was predicted using the I-sites library. The peptide fragment with the highest confidence prediction, FKKGERL, was synthesized and its structure characterized by proton NMR. The NMR parameters support the prediction that the peptide adopts a native-like diverging type II β-turn (Figure 2(b)). This is the first short peptide from an SH3 domain that has been demonstrated to adopt a native-like conformation in isolation.

I-sites predictions and the conformational preferences of isolated peptides

The diverging turn appears to be the lowest free energy conformation of the short peptide studied here. There are three factors that favor the diverging turn conformation for the peptide FKKGERL (Figure 2(b)): the hydrophobic interaction between the Phe at position 1 and the Leu at position 7, the Gly at position 4 allowing a positive phi angle, and the Glu at position 5 that forms a hydrogen bond with the backbone amide group at position 2. The pattern of sequence conservation in the I-sites profile shows each of these features, but it shows also features that can be explained only by negative design. Polar side-chains are conserved in positions 3 to 5 (Figure 2(a)); this prevents the formation of a stable amphipathic α-helix, since the latter requires conserved non-polar side-chains separated by three or four residues. Similarly, polar side-chains in positions 3 and 5 destabilize the amphipathic β-strand conformation, which prefers non-polar residues in those positions. Position 3 in the profile is always polar, but is never observed to be Asp or Asn. This may be negative design against the type I β-turn, which prefers Asp or Asn followed by Gly.

The lack of stable structure in the SH3 peptides studied by Serrano and colleagues (Viguera *et al.*, 1996) is consistent with the I-sites predictions. As illustrated in Figure 1(a), there is no high-confidence prediction spanning the five peptides that they studied. Two of the peptides contain portions of the diverging turn; however, neither contains the whole segment (the sequence MKKGDIL in the α-spectrin SH3 domain).

The confidence of a fragment prediction may be related to the fragment's free energy of folding in isolation. The confidence of I-sites predictions is defined as the fraction of peptide segments with a certain similarity score that have the predicted structure (Bystroff & Baker, 1998). For example, all seven residue peptide segments in the protein database were scored against the sequence profile for the diverging turn motif (Figure 2(a)), and 94% of all segments with a similarity score between 144 and 151 were found to have the diverging turn structure; therefore, a new sequence segment with a score of 147 has an estimated 94% probability of being a diverging turn. Because the segments in

the database from which the I-sites library was derived all differ in their respective global structures, the strong structural propensities must be due to internal contacts conserved within I-sites clusters. Clearly, the relation of confidence to equilibrium concentration, and through this to free energy, is a crude one. Nonetheless, the evidence suggests that the relation holds qualitatively, at least for sequence segments with high confidence scores.

The studies by Dyson *et al.* (1992) on short peptides comprising the entire length of the β-sandwich protein plastocyanin showed that only a small number of peptides had any tendency to form structure in isolation. Interestingly, the peptide with the strongest structural tendencies was a diverging β-turn similar to that studied here. To further explore the relationship between I-sites predictions and peptide conformational preferences, I-sites predictions were made for the plastocyanin sequence family. The three regions found previously to have some native structural tendencies in isolation were all contained within correct I-sites structure predictions with a confidence of 0.60 or greater.

We present several more I-sites predictions (Table 2) to be confirmed or refuted by future experimental evidence. These short sequences were chosen mostly from small proteins, some of which are presently under intensive experimental study. They are predicted to adopt a significant amount of native structure in isolation. Structure-blind I-sites predictions can be made automatically *via* http://ganesh.bchem.washington.edu/~bystroff/Isites/.

Role of diverging turn in protein folding

It is interesting to compare our results with recent structural information obtained on unfolded states of the drkN SH3 domain under both folding conditions (U_{exch}) and denaturing conditions (U_{Gnd}; Zhang & Forman-Kay, 1997). The most marked differences between the U_{exch} and U_{Gnd} states are located in the segment $_{19}$FRKTQILKIL$_{28}$, which corresponds to $_{26}$FKKGERLQIV$_{35}$ in the src SH3 domain (the diverging turn segment is underlined). A turn conformation was populated around $_{19}$FRKT$_{22}$ in U_{exch}, but not in U_{gnd}, and residues in $_{23}$QILKIL$_{28}$ exhibited severe line-broadening in the U_{exch} state that disappeared upon addition of denaturant, suggesting that this portion of the chain adopts a structure that is in intermediate exchange with other conformational substates. By analogy to our results, the conformational substates populated by this portion of the drk U_{exch} may include a similar type of diverging turn conformation, which may be less stable in drk than in src because of the replacement of the glycine residue by threonine (Figure 2(a)).

The characteristic features of the diverging turn sequence motif are highly conserved in all known SH3 domains (Koyama *et al.*, 1993). Each of the other hairpin turns in the molecule shows evolutionary variability in both sequence and length. This portion of the structure is less variable among the three-dimensional structures of SH3 domains solved to date than any other segments outside of the hydrophobic core (Guruprasad *et al.*, 1995). The strong sequence and structural conservation within the diverging turn along with the drk denatured state and the src peptide studies may indicate an important role for the turn in the folding process.

It is interesting that the turn in the SH3 domain with the strongest propensity to adopt structure in isolation is not a tight β-hairpin; instead, the turn connects strands that do not share backbone hydrogen bonds (the earlier peptide studies on α-spectrin SH3 focused on the β-hairpins). The SH3 domain may be viewed as two orthogonally packed β-sheets, where the diverging β-turn is one of the two transitions between the sheets in the protein. Formation of the diverging β-turn conformation early in folding could play an important role in establishing the topology of the protein by preventing inappropriate formation of a β-hairpin and promoting proper packing of hydrophobic side-chains between the diverging strands. The

Table 2. Blind predictions of folding initiation sites

PDB code	Seq. numbers	Sequence	I-sites motif/native structure	C_f^a
1aaj	37-45	VKVGDTVTWI	Diverging turn	1.00
1aaj	72-80	MKKEQAYSL	Diverging turn	0.85
1edt	77-87	RPLQQQGIKVL	Helix C-cap, type 1	1.00
1edt	50-57	YDTGTKTA	Ser beta-hairpin	0.80
1htp	106-112	TSPDELE	Helix N-cap	1.00
1htp	115-121	LGAKEYTKFC	Helix N-cap	1.00
1ifc	14-25	YEKFMEKMGINV	Helix C-cap, type 2	1.00
3aah	393A-401A	YDPESRTLY	Ser-hairpin	1.00
1bgl	576A-585A	SLIKYDENGNPW	Type I hairpin	1.00
1nhp	63-74	GEKMESRGVNVF	Helix C-cap, type 2	1.00
1trk	316A-326A	FSEYQKKFPEL	Pro helix C-cap	0.90

Columns 1 and 2 give the PDB code and residue names, and column 3 gives the sequences of segments predicted to adopt native structure in isolation. The I-sites sequence-structure motif is listed in column 4, and column 5 is the confidence of the prediction. These fragments were selected at random from a large set of predictions. A prediction server is available at the web site http://ganesh.bchem.washington.edu/~bystroff/Isites/.

observation of a structured diverging turn peptide in plastocyanin suggests that such a role may be a common feature of the folding of β-sheet proteins. Our results show that the diverging turn in src-SH3 is stable in isolation and, because all of the interactions are local, it undoubtedly forms very rapidly. This is consistent with kinetic studies of src SH3 mutants (Grantcharova *et al.*, 1998), which suggest that the diverging turn and the following strand come together with the distal loop hairpin in the folding transition state.

An attractive feature of the protein folding problem is that it is amenable to both computational and experimental approaches. The work described here illustrates that a combination of these approaches can provide significantly more insight than either one alone. *Ab initio* structure prediction methods even in their current imperfect state can generate hypotheses to guide experimental studies of the folding process.

Materials and Methods

Prediction of peptide conformation

A sequence profile (Gribskov *et al.*, 1990) was constructed from an alignment of SH3 domain sequences in the HSPROT database, aligned initially *via* the PHD server (Rost *et al.*, 1994) then modified to agree with earlier structural alignments (Feng *et al.*, 1995; Guruprasad *et al.*, 1995). All subfragments of the profile were scored against all motifs in the library as described elsewhere (Bystroff & Baker, 1998). Scores were translated into confidence values using the results of cross-validation studies on a large non-redundant database of proteins of known structure; the confidence of a prediction is simply the probability that the prediction is correct.

NMR examination of the peptide conformation

The sequence Phe-Lys-Lys-Gly-Glu-Arg-Leu was synthesized and purified by Research Genetics Co. (Huntsville, AL), and the molecular mass was confirmed by MALDI-MS. Two forms of the peptide, one with free amino and carboxyl groups at the N and C termini, the other with an acetylated N terminus and an amidated C terminus, were investigated by NMR. NMR samples were prepared by dissolving ~10 mg of peptide in 0.45 ml of 50 mM sodium phosphate buffer (90% H_2O/10% ^2H_2O). The pH of the samples were adjusted using diluted NaOH or HCl solutions. TSP was added to the NMR samples to a final concentration of 0.5 mM for chemical shift referencing.

NMR spectra were acquired on a Bruker DMX500 spectrometer at 12°C unless otherwise specified. ^{1}H-TOCSY spectra (Bax & Davis, 1985a) were collected using spectral widths of 6250 Hz in both dimensions and 1024 × 600 complex points. ^{1}H-ROESY spectra (Bax & Davis, 1985b) were collected with a spin-lock field strength of 8.62 kHz, a spectral width of 6250 Hz in both dimensions and 1024 × 600 complex points. The mixing time for the TOCSY and ROESY experiments were 50 ms and 250 ms, respectively. Data were apodized with a squared sine bell in both dimensions, and zero-filled to give 1 K × 1 K complex spectra. Water supression was achieved with the watergate pulse sequence (Sklenar *et al.*, 1993). The recycle delay was 2.2 s for all the experiments. All the data processing was performed on an SGI workstation using the program NMRpipe (Delaglio *et al.*, 1995).

The $^3J_{NH_\alpha}$ coupling constants were obtained directly from the resolved amide proton resonances in the 1D spectrum collected with a digital resolution of 0.43 Hz/point. The temperature coefficients of the amide protons were obtained from linear fits of the chemical shift data from 1D spectra acquired at 9, 12, 15, 18, 21, 24, 27, 30, 33 and 36°C.

Acknowledgements

We thank David Shortle, Patricia Campbell, Kevin Plaxco and Kim Simons for helpful comments on the manuscript. This work was supported by young investigator awards to D.B. from the NSF and Packard foundation.

References

Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1995). Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J. Mol. Biol.* **252**, 460–471.

Avbelj, F. & Moult, J. (1995). Determination of the conformation of folding initiation sites in proteins by computer simulation. *Proteins: Struct. Funct. Genet.* **23**, 129–141.

Bax, A. & Davis, D. G. (1985a). MLEV-17-based two-dimensional homonuclear magnetization transfer spectroscopy. *J. Magn. Reson.* **65**, 355–360.

Bax, A. & Davis, D. G. (1985b). Practical aspects of two-dimensional transverse NOE spectroscopy. *J. Magn. Reson.* **63**, 207–213.

Bystroff, C. & Baker, D. (1997). Blind ab initio local structure predictions using a library of sequence-structure motifs. *Proteins: Struct. Funct. Genet.* **(Suppl. 1)**, 167–171.

Bystroff, C. & Baker, D. (1998). Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.* **281**, 565–577.

Campbell, A. P., McInnes, C., Hodges, R. S. & Sykes, B. D. (1995). Comparison of NMR solution structures of the receptor binding domains of Pseudomonas aeruginosa pili strains PAO, KB7, and PAK: implications for receptor binding and synthetic vaccine design. *Biochemistry*, **34**, 16255–16268.

Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfiefer, J. & Bax, A. (1995). NMRpipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biol. NMR*, **6**, 277–293.

Doyle, R., Simons, K., Qian, H. & Baker, D. (1997). Local interactions and the optimization of protein folding. *Proteins: Struct. Funct. Genet.* **29**, 282–291.

Dyson, H. J., Rance, M., Houghten, R. H., Lerner, R. A. & Wright, P. E. (1988). Sequence requirements for the formation of a reverse turn. *J. Mol. Biol.* **201**, 161–200.

Dyson, H. J., Sayre, J. R., Merutka, G., Shin, H. C., Lerner, R. A. & Wright, P. E. (1992). Folding of peptide fragments comprising the complete sequence of proteins. Models for initiation of protein folding. II. Plastocyanin. *J. Mol. Biol.* **226**, 819–835.

Feng, S., Kasahara, C., Rickles, R. J. & Schreiber, S. L. (1995). Specific interactions outside the proline-rich core of two classes of Src homology 3 ligands. *Proc. Natl Acad. Sci. USA*, **92**, 12408–12415.

Fersht, A. R. (1995). Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc. Natl Acad. Sci. USA*, **92**, 10869–10873.

Grantcharova, V. P., Riddle, D. S., Santiago, J. V. & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nature Struct. Biol.* **5**, 714–720.

Gribskov, M., Luthy, R. & Eisenberg, D. (1990). Profile analysis. *Methods Enzymol.* **183**, 146–159.

Guruprasad, L., Dhanaraj, V., Timm, D., Blundell, T. L., Gout, I. & Waterfield, M. D. (1995). The crystal structure of the N-terminal SH3 domain of Grb 2. *J. Mol. Biol.* **248**, 856–866.

Koyama, S., Yu, H., Dalgarno, D. C., Shin, T. B., Zydowsky, L. D. & Schreiber, S. L. (1993). Structure of the PI3 K SH3 domain and analysis of the SH3 family. *Cell*, **72**, 945–952.

Merritt, E. A. & Murphy, M. E. P. (1994). Raster3D version 2.0. A program for photorealistic molecular graphics. *Acta Crystallog. Sect. D*, **6**, 869–873.

Muñoz, V. & Serrano, L. (1996). Local versus nonlocal interactions in protein folding and stability – an experimentalist's point of view. *Folding Des.* **1**, R71–R77.

Riddle, D. S., Santiago, J. V., Bray-Hall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q. & Baker, D. (1997). Functional rapidly folding proteins from simplified amino acid sequences. *Nature Struct. Biol.* **4**, 805–809.

Rose, G. D., Gierasch, L. M. & Smith, J. A. (1985). Turns in peptides and proteins. *Advan. Protein Chem.* **37**, 1–106.

Rost, B., Sander, C. & Schneider, R. (1994). PHD–an automatic mail server for protein secondary structure prediction. *Comput. Appl. Biosci.* **10**, 53–60.

Sklenar, V., Piotto, M., Leppik, R. & Saudek, V. (1993). Gradient-tailored water suppression for 1H-15N HSQC experiments optimized to retain full sensitivity. *J. Magn. Reson. Ser. A*, **102**, 241–245.

Smith, L. J., Bolin, K. A., Schwalbe, H., MacArthur, M. W., Thornton, J. M. & Dobson, C. M. (1996). Analysis of main chain torsion angles in proteins: Prediction of NMR coupling constants for native and random coil conformations. *J. Mol. Biol.* **255**, 494–506.

Unger, R. & Moult, J. (1996). Local interactions dominate folding in a simple protein model. *J. Mol. Biol.* **259**, 988–994.

Viguera, A. R., Jiménez, M. A., Rico, M. & Serrano, L. (1996). Conformational analysis of peptides corresponding to beta-hairpins and a beta-sheet that represent the entire sequence of the alpha-spectrin SH3 domain. *J. Mol. Biol.* **255**, 507–521.

Wang, Y. & Shortle, D. (1997). Residual helical and turn structure in the denatured state of staphylococcal nuclease: analysis of peptide fragments. *Folding Des.* **2**, 93–100.

Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids*, John Wiley & Sons, Inc., New York.

Xu, W., Harrison, S. C. & Eck, M. J. (1997). Three dimensional structure of the tyrosine kinase C-src. *Nature*, **385**, 595–602.

Zhang, O. & Forman-Kay, J. D. (1997). NMR studies of unfolded states of an SH3 domain in aqueous solution and denaturing conditions. *Biochemistry*, **36**, 3959–3970.

Edited by P. E. Wright

(Received 10 April 1998; received in revised form 2 July 1998; accepted 7 July 1998)

1. Mirny, L.A., Abkevich, V.I. & Shakhnovich, E.I. *Proc. Natl. Acad. Sci. USA* **95**, 4976–4981 (1998).
2. Michnick, S.W. & Shakhnovich, E. *Folding & Design* **3**, 239–251 (1998).
3. Plaxco, K.W., Simons, K.T. & Baker, D. *J. Mol. Biol.* **277**, 985–994 (1998).
4. Shakhnovich, E. *Folding & Design* **3**, R108–R111 (1998).
5. Thirumalai, D. & Klimov, D.K. *Folding & Design* **3**, R112–R118. (1998).
6. Martinez, J.C., Pisabarro, M.T. & Serrano, L. *Nature Struct. Biol.* **5**, 721–729 (1998).
7. Musacchio, A., Noble, M., Pauptit, R., Wierenga, R. & Saraste, M. *Nature* **359**, 851–855 (1992).
8. Blanco, F.J., Ortiz, A.R. & Serrano, L. *J. Biomol. NMR* **9**, 347–357 (1997).
9. Viguera, A.R., Martinez, J.C., Filimonov, V.V., Mateo, P.L. & Serrano, L. *Biochemistry* **33**, 2142–2150 (1994).
10. Viguera, A.R., Blanco, F.J. & Serrano, L. *J. Mol. Biol.* **247**, 670–681 (1995).
11. Viguera, A.R., Serrano, L. & Wilmanns, M. *Nature Struct. Biol.* **3**, 874–880 (1996).
12. Viguera, A.R. & Serrano, L. *Nature Struct. Biol.* **4**, 939–946 (1997).
13. Grantcharova, V.P., Riddle, D.S., Santiago, J.V. & Baker, D. *Nature Struct. Biol.* **5**, 714–720 (1998).
14. Fersht, A.R. *Curr. Opin. Struct. Biol.* **5**, 79–84 (1995).
15. Fersht, A.R., Itzhaki, L.S., elMasry, N.F., Matthews, J.M. & Otzen, D.E. *Proc. Natl. Acad. Sci. USA* **91**, 10426–10429 (1994).
16. Riddle D.S. *et al. Nature Struct. Biol.* **6**, 1016–1024 (1999)
17. Chiti, F., *et al. Nature Struct. Biol.* **6**, 1005–1009 (1999).
18. Villegas, V., Martinez, J.C., Avilés, F.X. & Serrano, L. *J. Mol. Biol.* **283**, 1027–36 (1998).
19. Kunkel, T.A.. *Proc. Natl. Acad. Sci. USA* **82**, 488–492 (1985).
20. Gill, S.C. & Hippel, P.H. *Anal. Biochem.* **182**, 319–326 (1989).
21. Prieto, J., Wilmans, M., Jimenez, M.A., Rico, M. & Serrano, L. *J. Mol. Biol.* **268**, 760–778 (1997).
22. Johnson, C.M. & Fersht, A.R. *Biochemistry* **34**, 6795–6804 (1995).

Experiment and theory highlight role of native state topology in SH3 folding

David S. Riddle[1-3], Viara P. Grantcharova[1,2], Jed V. Santiago[2], Eric Alm[2], Ingo Ruczinski[2] and David Baker[2]

[1]*These authors contributed equally to this work.* [2]*Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA.* [3]*Present address: Department of Immunology, Mayo Clinic, Rochester, Minnesota 55904, USA.*

We use a combination of experiments, computer simulations and simple model calculations to characterize, first, the folding transition state ensemble of the src SH3 domain, and second, the features of the protein that determine its folding mechanism. Kinetic analysis of mutations at 52 of the 57 residues in the src SH3 domain revealed that the transition state ensemble is even more polarized than suspected earlier: no single alanine substitution in the N-terminal 15 residues or the C-terminal 9 residues has more than a two-fold effect on the folding rate, while such substitutions at 15 sites in the central three-stranded β-sheet cause significant decreases in the folding rate. Molecular dynamics (MD) unfolding simulations and *ab initio* folding simulations on the src SH3 domain exhibit a hierarchy of folding similar to that observed in the experiments. The similarity in folding mechanism of different SH3 domains and the similar hierarchy of structure formation observed in the experiments and the simulations can be largely accounted for by a simple native state topology-based model of protein folding energy landscapes.

Three independent lines of investigation suggest that protein folding rates and mechanisms are largely determined by native state topology[1]. First, dramatic changes in amino acid sequence, produced either in the laboratory[2,3] or by the evolutionary process[4], that do not alter the overall topology of a protein usually have relatively little effect on protein folding rates. Second, comparison of the consequences of mutations on folding kinetics in distantly related homologs suggests that folding transition state structure is conserved despite differences in amino acid sequence and stability[5,6]. Third, the folding rates of small proteins are strongly correlated with a property of the native state topology: the average sequence separation between residues that make contacts in the three-dimensional structure (the contact order)[7]. The influence of native state topology on protein folding rates and mechanisms is a consequence of the relatively large entropic cost of forming nonlocal interactions early in folding: simple topologies with mostly local interactions are more readily formed than those with many nonlocal interactions, and for a given topology, local interactions are more likely to be formed early in folding than nonlocal interactions.

SH3 domains are an ideal system to investigate how topology determines folding mechanisms. Over 400 different naturally occurring SH3 domain sequences have been identified, more than 10 high-resolution structures have been determined, and the stability and folding kinetics of a number of these proteins have been characterized[6,8–14]. We had found that many of the residues conserved in a phage display selection for simplified src SH3 domain variants[2] played an important role in determining the folding mechanism[5]. Kinetic analysis of mutations in 20 of the 57 positions in the protein suggested that the distribution of structure in the transition state ensemble was localized to one portion of the molecule, and that the folding transition state of the src SH3 domain resembled that of the α-spectrin SH3 domain, which has an almost identical topology but only 36% sequence identity[6]. Though suggestive, with less than half of the residues accounted for, these results did not thoroughly characterize the transition state ensemble of the protein, and did not provide an explanation for the similarity in the src and spectrin SH3 transition states.

In this paper we present a combination of experiments, computer simulations, and simple model calculations aimed at detailed characterization of the src SH3 transition state and its structural origins. The experiments fully map out the transition state ensemble by probing the kinetic consequences of mutations of every residue that makes appreciable interactions in the native state. The computer simulation studies assess the robustness of the hierarchy of structure formation to the numerous approximations and likely inaccuracies in computational models of folding. Finally, the simple model calculations probe the topological features that determine the way SH3 domains fold. Our results provide perhaps the most comprehensive picture of the rate-limiting step in folding of an all β-sheet protein available to date.

Experimental studies

The SH3 domain is a 57-residue globular protein that consists of two antiparallel β-sheets orthogonally packed to form a single hydrophobic core (Fig. 1). Here we describe the effects of mutations of all residues more than 10% buried in the native

structure (52 of 57 residues in the protein) on the rates of folding and unfolding, and the picture of the folding transition state that emerges from these data.

The method we employ was pioneered by Fersht and coworkers[15] and has emerged as the predominant experimental procedure for the detailed characterization of folding transition states[16–20]. The extent to which a residue's interactions are formed in the transition state is summarized by the Φ_f value ($\Delta\Delta G_{u-\ddagger}/\Delta\Delta G_{u-f}$), which is the change in the free energy of the transition state brought about by mutation of the residue normalized by the change in overall stability[15]. A Φ_f value of 1 indicates that all of a residue's interactions are formed in the transition state, whereas a Φ_f value of 0 means that the residue does not make stabilizing interactions in the transition state. Intermediate Φ_f values indicate partially formed interactions or interactions formed in a fraction of the transition state ensemble; the relationship between the actual Φ_f value and the extent of structure formation is not necessarily linear. As emphasized by Fersht and coworkers[16], the most straightforward class of mutations to interpret are those that remove a small number of methyl groups, such as isoleucine to valine, alanine to glycine, and valine to alanine, as these are least likely to change the folding mechanism and the structure of the folded and unfolded states. In this study, we have also mutated polar residues to alanine to examine the role of polar interactions and hydrogen bonds in the transition state, and have substituted glycine residues with alanine to probe turn formation in the transition state. To guard against possible artifacts due to changes in denatured state structure and/or folding mechanism, we draw conclusions only from results that are consistent among a number of neighboring residues.

To facilitate presentation of our results, we have divided the src SH3 domain into five structural regions and discuss them in order of increasing importance in the folding transition state.

N- and C-terminal strands (strands 1 and 5) and 3_{10}-helix. The N- and C-termini of the SH3 domain come together to form an antiparallel β-sheet stabilized by nonlocal side chain–side chain interactions (Fig. 1 and Table 1). A short 3_{10}-helix (PSNY, residues 57–60) precedes strand 5 and is responsible for the 90° transition from one sheet to the other. It is remarkable that almost all mutations in this region (12 out of 14) either exclusively affect the unfolding rate or do not change protein stability (Fig. 2*a*,*b*).

The extremely low Φ_f values (Fig. 1, Table 2) suggest that the N- and C-termini are largely unstructured in the transition state ensemble.

RT Loop. Residues 14–25 (YDYESRTETDLS) form the large, relatively disordered RT loop (Fig. 1), which is functionally important for binding proline-rich peptides. The crystal structure of the src SH3 domain reveals a small stretch of regular β-sheet pairing within the loop, as well as quite a few intraloop hydrogen bonds involving the side chains of D15, S18, D23 and S25 (Table 1). Hydrogen/deuterium (HD) exchange experiments[8] indicate, however, that this part of the molecule is flexible. As with the N- and C-termini, almost all mutations (eight out of nine) in the RT loop have Φ_f values close to 0 (Fig. 2*c*, Table 2). L24A is the only mutation in the RT loop that lowers k_f, but its predominant effect is still on k_u. The RT loop and the N- and C-termini are clearly the parts of the SH3 domain that are least structured in the transition state (Fig. 1).

Diverging type II β-turn. The transition from the RT loop to the central three-stranded sheet formed by the n-src loop and the distal loop β-hairpin is made by the diverging turn (FKKGERLQ, residues 26–33) (Fig. 1). It is stabilized by hydrophobic contacts between the central core residues, F26 and L32, and a hydrogen bond between the side chain carboxyl of E30 and the backbone amine of K27.

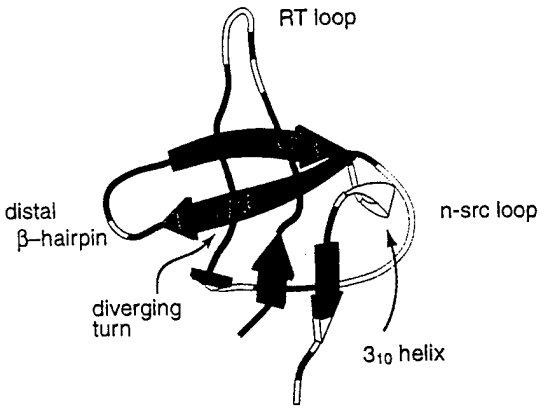


Fig. 1 Structure[35] of the src SH3 domain colored by Φ_f value from red (1) to blue (0). Residues colored in white were either not mutated or the mutation did not affect ΔG significantly. Residues colored in yellow increased or decreased both k_f and k_u, suggesting that these mutations affect the transition state more than the native state. Φ_f values were calculated as described in the Methods. The image was created using Molscript[36].

All of the structurally important residues in this region have intermediate Φ_f values (Fig. 2*e*, Table 2). NMR studies[21] of an isolated peptide with the sequence FKKGERL suggest that the diverging turn conformation is partially populated in the denatured state. Thus, the interactions made by the diverging turn residues in the transition state may be greater than indicated by the Φ_f values, since the reference state (the denatured state) is already partially ordered. Recent double-mutant experiments (V.G. and D.B., unpublished results) suggest that the additional interactions made by the diverging turn in the transition state include a nonlocal hydrogen bond network involving E30 in the diverging turn and S47 and T50 in the distal β-hairpin. The partial Φ_f values of the core residues F26 and L32 suggest that these residues also make some of their interactions with hydrophobic residues in the distal loop β-hairpin in the transition state.

N-src loop. The n-src loop (IVNNTEGDWW, residues 34–43) (Fig. 1) has an unusual shape: the two end residues, I34 and W43, are part of the hydrophobic core whereas the intervening sequence forms a large, almost rectangular turn around W43. W42 is only peripherally associated with the core and together with W43 lines the peptide binding site. There is limited local hydrogen bonding within the n-src loop, and two nonlocal hydrogen bonds connect it to the 3_{10}-helix (Table 1).

The large number of mutations with unusual kinetic consequences suggests that this region may adopt nonnative conformations in the transition state (Fig. 2*d*, Table 2). I34 is a central hydrophobic core residue with many neighbors (Table 1), yet neither I34A nor I34V affect stability significantly (ΔΔG -0.33 and -0.09 kcal mol^{-1}, respectively). Kinetic analysis shows that the two I34 mutants slow both the folding and unfolding rates simultaneously, suggesting that the mutations destabilize the transition state more than the native or denatured states. I34 appears to be critical for core formation during folding, but strained in the native state because of slight overpacking of

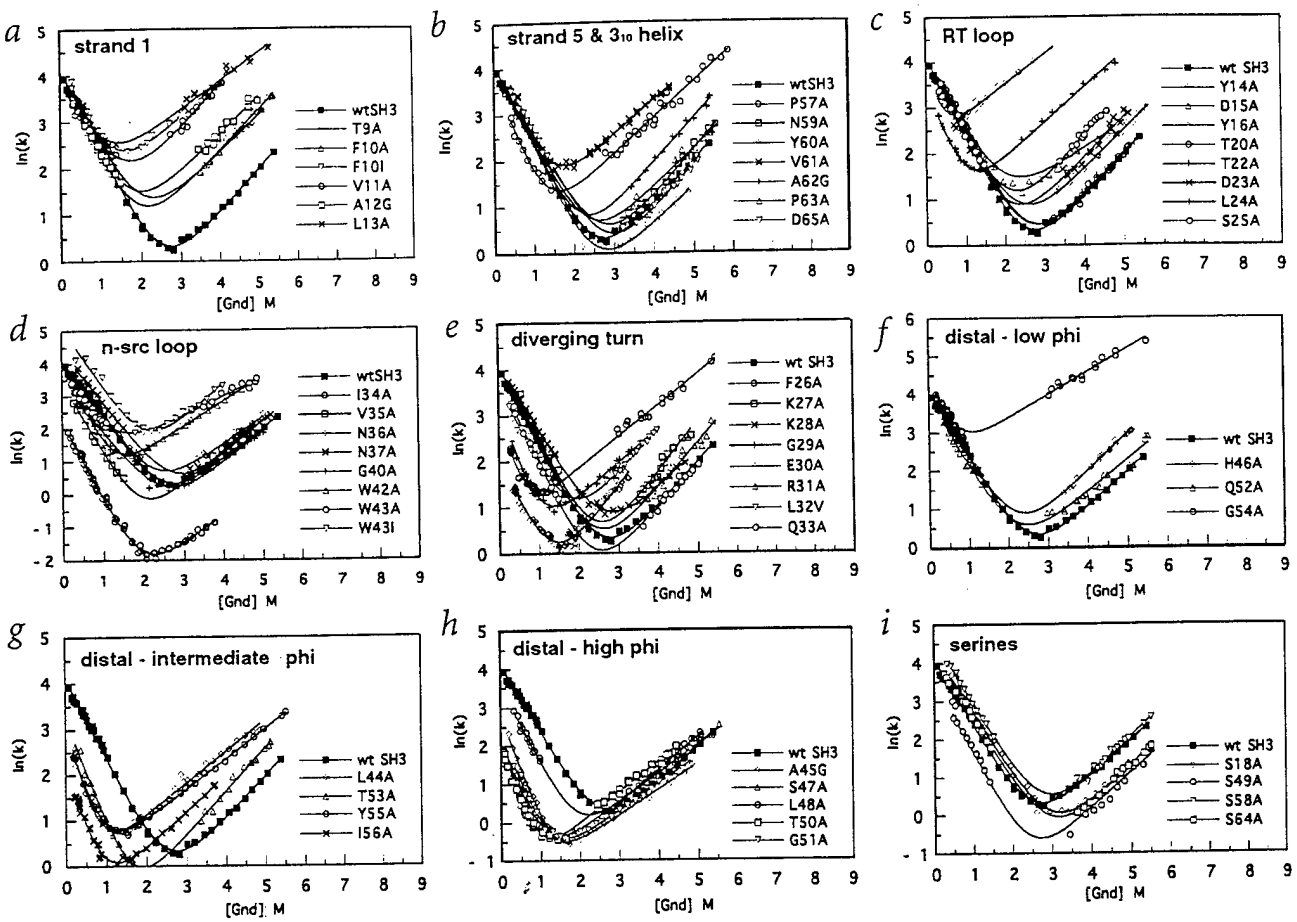


Fig. 2 Dependence of the rate of folding and unfolding on the denaturant concentration for all the mutants grouped into structural regions (*a–h*) as shown in Fig. 1. *i*, Serine to alanine substitutions with unusual behavior. The data for the wild type (wt) protein (■) is shown in all panels for comparison. The solid lines represent the fits to the experimental data.

the hydrophobic core in the native src SH3 domain. Overpacking is most likely due to burial of the bulky W43 in the native state, but not in the transition state ($\Phi_f = 0.15$ for the W43A mutation). On the solvent-exposed side, V35A is the only mutation for which an unambiguous Φ_f value can be calculated (0.77); its interaction with L44 in the distal β-hairpin appears to be partially formed in the transition state. The N37 side chain appears to make unfavorable interactions in the transition state as the N37A mutation speeds both folding and unfolding. Chain reversal at the tip of the n-src loop appears to be important at the transition state as the G40A mutation, which stiffens the chain, slows both k_f and k_u. As mutations in I34, N37 and G40 appear to selectively stabilize or destabilize the transition state relative to the native and unfolded states, it seems likely that the residues in the n-src loop are ordered at the transition state, but in a nonnative conformation (perhaps a tight hairpin, rather than the distorted loop present in the native state).

Distal β-hairpin. Strands 3 (LAHS, residues 44–47) and 4 (RTGYI, residues 52–56) form the distal β-hairpin, the most regular element of secondary structure in the SH3 domain (Fig. 1). They are connected by a tight type I β-turn and stabilized by numerous backbone and side chain hydrogen bonds, including an extensive network of hydrogen bonds among the turn residues S47 and T50 and the peptide backbone (Table 1).

Mutations throughout the distal β-hairpin can be grouped into three categories based on their effects on kinetics (Table 2). Mutations with Φ_f values of 0 (Fig. 2*f*) include H46A and Q52A (both exposed polar residues) and G54A. Among the residues with intermediate Φ_f values (Fig. 2*g*, Table 2) L44, T53 and Y55 interact at the solvent-exposed side of the hairpin and appear to be only partially associated at the transition state. I56, on the other hand, is an integral part of the hydrophobic core and intimately involved in the transition state as judged by the large decrease in k_f upon mutation to alanine; it does, however, make additional interactions after the transition state as well ($\Phi_f = 0.71$). Mutations with Φ_f values of 1 are clustered around the turn (S47A, L48A, T50A and G51A) or are part of the hydrophobic core (A45G) (Fig. 2*h*), suggesting that the β-turn is fully formed in the transition state and the center of the β-hairpin is associated with the hydrophobic core. As mentioned earlier, residues S47 and T50 also make nonlocal interactions with the diverging turn. S49 (Fig. 2*i*) might also take part in this hydrogen bond network at the transition state as the S49A mutation decreases k_f; the decrease in k_u brought about by the mutation may be due to partial burial of the -OH group in the native state without a suitable hydrogen bonding partner.

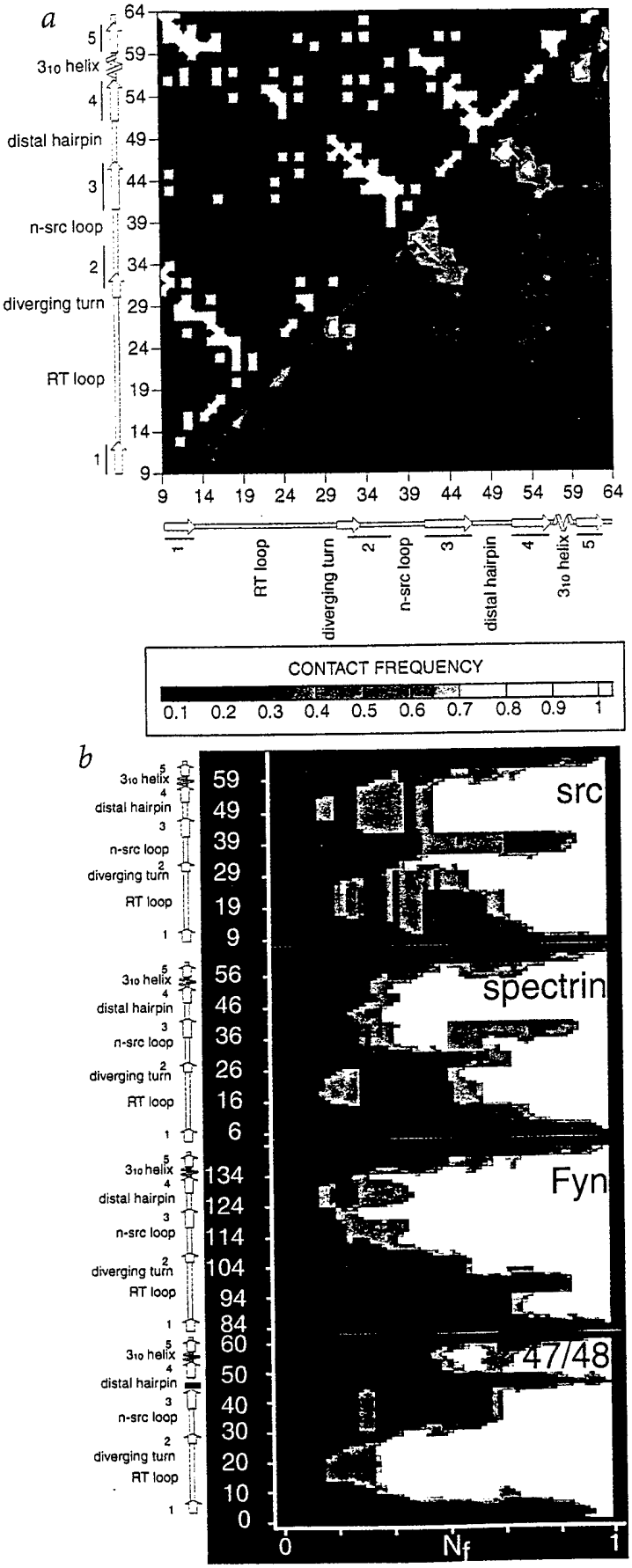


Fig. 3 Theoretical analysis of SH3 folding. ***a***, *Ab initio* simulation of src SH3 folding using ROSETTA. The folding of the src SH3 domain was simulated using ROSETTA as described in Methods. All SH3 domain structures were removed from the data base of short fragments used for building up conformations. A total of 500 independent simulations were carried out, and all conformations from the 20 trajectories that produced structures within 4.5 Å r.m.s.d. of the native structure were combined to calculate the frequency of side chain–side chain contacts for each pair of residues in the protein (lower right triangle; color scheme is shown below the figure). For comparison, the contact distribution in the native structure is shown in the upper left. ***b***, Hierarchy of SH3 domain folding in model calculations based on native state topology. Calculations were performed on the src, spectrin and fyn SH3 domains and the 47–48 circular permutant of the spectrin SH3 domain in which the distal hairpin has been cut. The reaction coordinate, N_f, is the fraction of ordered residues ($N_f = 0$ is the fully unfolded state and $N_f = 1$ is the fully folded state). The y-axis indicates position along the sequence. All configurations of the system were enumerated, and the Boltzmann averaged frequency of ordering of each residue, as a function of N_f, is indicated by the color (black-blue, 0–0.25; blue-magenta, 0.25–0.50; magenta-red, 0.50–0.75; red-yellow, 0.75–0.88; and yellow-white, 0.88–1.00). The top panel was shown in Alm and Baker[31]. It is important to note that segments of the protein not contiguous along the sequence still interact in the model if contacting in the three-dimensional structure, for example in the top panel, the high population of the diverging turn/strand 2 and the distal loop β-hairpin at $N_f = 0.6$ indicates that more surface area is buried within and between these structural elements than within any other substructure with the same number of residues ordered in the protein.

Putting the pieces together, the following picture of the transition state of the src SH3 domain emerges (Fig. 1). The distal β-hairpin is the most ordered structural element in the transition state. The diverging turn and strand 2 are partially ordered and interact with residues in the distal β-hairpin, and this effectively constrains the n–src loop and specifies the three-stranded topology of the central β-sheet in the protein. The clustering of mutations that selectively stabilize or destabilize the transition state in the vicinity of the n-src loop (Fig. 1, yellow) suggests that the loop may have a nonnative configuration in the transition state. In contrast, the two terminal strands, the RT loop and the 3_{10}-helix are mostly unstructured and contribute few stabilizing interactions in the transition state.

Because of the complexities associated with interpreting any one mutation, consistency within a large set of mutations is critical for constructing a plausible picture of structure in the transition state. The segment of sequence between residues 26 and 58 contains only 28 of the 43 positions for which mutations significantly affected the rate of folding and/or unfolding, but this segment contains 25 of the 27 positions with either (i) Φ_f values greater than 0.15 or (ii) mutations that selectively stabilize or destabilize the transition state (Table 2). The probability of such a partitioning if the observed Φ_f values were randomly distributed in the sequence is 1 in 758,000.

Computer simulations

To further elucidate the hierarchy of structure formation in the SH3 domain, we compare our experimental findings with the results from two complementary computational methods: recently published molecular dynamics (MD) simulations of src SH3 domain

Table 1 Interactions in the native state of src SH3 domain[1]

	Residue	Burial %	Hydrophobic interactions[2]	sm H bonds[3]	ss H bonds[4]
	T9	13	11, 31, 48, 64		33
β1	F10	76	32, 34, 37, 43, 61, 63		
	V11	64	9, 13, 31, 64		
	A12	100	26, 30, 32, 61		
	L13	54	11, 14, 62		
	Y14	44	13, 28, 60		
	D15	48	17, 28	28	
	Y16	85	19, 23, 26, 42, 56, 57, 60		23
	E17	25	15		
RT	S18	87	23	25	
loop	R19	8	16, 20, 42	23	
	T20	39	19, 23	22, 23	
	E21	7	22		
	T22	26	21, 55	20	
	D23	79	16, 18, 20, 42, 55	19, 20	16
	L24	69	25, 26, 32, 45, 47, 52, 56		
	S25	51	24	18	
	F26	90	12, 16, 24, 30, 32, 56, 57, 60, 61		
	K27	34	30, 50	30	
diverging	K28	30	14, 15	15	
turn	G29	22			
	E30	53	12, 26, 27, 32, 47, 49, 50	27	47
	R31	21	9, 11, 64		
β2	L32	94	10, 12, 24, 26, 30, 34, 45, 47, 56, 61		
	Q33	46	35, 46, 48		9
	I34	77	10, 32, 37, 43, 45, 56, 61		
	V35	51	33, 36		
	N36	28	35, 38		
n-src	N37	53	10, 34, 43		
loop	T38	7	36, 43	43	
	E39	21			
	G40	33		58	
	D41	43	42		
	W42	50	16, 19, 23, 41, 55, 57		
	W43	92	10, 34, 37, 38, 56, 58, 61	38	
β3	L44	83	53, 55		
	A45	99	24, 32, 34, 56		
	H46	68	33, 48, 53		
	S47	79	24, 30, 32, 52	49, 50	30
	L48	36	9, 33, 46		
	S49	12	30, 50	47	
distal	T50	21	27, 30, 49, 52	47	
hairpin	G51	54			
	Q52	44	24, 47, 50		
	T53	53	44, 46		
β4	G54	94			
	Y55	62	22, 23, 42, 44		
	I56	99	16, 24, 26, 32, 34, 43, 45, 57, 61		
	P57	96	16, 26, 42, 56, 60		
3_{10}	S58	87	43	40	
helix	N59	16	60		
	Y60	63	14, 16, 26, 57, 59, 61		
	V61	95	10, 12, 26, 32, 34, 43, 56, 60		
β5	A62	77	13, 63		
	P63	34	10, 62		
	S64	0	9, 11, 31, 65		
	D65	0	64		

[1]The crystal structure of the src tyrosine kinase[35] was used as a model for the native state of the SH3 domain (1fmk).
[2]Hydrophobic interactions were defined using the Voronoi procedure[22].
[3]Main chain–side chain hydrogen bonding.
[4]Side chain–side chain hydrogen bonding.

unfolding[22], and folding simulations of the src SH3 domain using our *ab initio* folding method, ROSETTA, which recently showed considerable promise in structure prediction in the CASP3 experiment[23] (folding of the SH3 domain starting with an unfolded polypeptide is computationally prohibitive using MD; ROSETTA achieves the vast speed-up necessary by simplifying both the conformational search strategy and the potential function). The chain representations, potentials and conformational sampling methods used by the two approaches are radically different; any common features observed in the two simulations are thus likely to reflect properties of the overall fold rather than specific residue–residue interactions.

ROSETTA uses local structural information from the protein data base and a simplified potential function to fold amino acid sequences to compact protein-like structures[23–25] (see Methods). Even for a small protein such as the SH3 domain, only a fraction of ROSETTA trajectories pass through native-like conformations. Inspection of individual successful trajectories suggested that the order of events in folding was quite similar to that observed experimentally. To get a more quantitative picture of the conformations sampled, we identified the substructures populated most frequently in 20 successful folding trajectories that produced structures within 4.5 Å r.m.s.d. (on Cα) of the native state. The occupancy of all side chain–side chain contacts (both native and nonnative) was averaged over all conformations in the 20 trajectories (Fig. 3*a*). Interactions formed early in the trajectory and persisting throughout have high occupancy in Fig. 3*a*, whereas contacts formed late have low occupancy. Thus, while this analysis does not single out the transition state ensemble (this could potentially be done using the p_{fold} method[26], but would be extremely computationally expensive), it provides information about the overall hierarchy to folding in the simulations. As is evident in Fig. 3*a*, the distal β-hairpin, the n-src loop and the diverging turn are highly populated during the simulations, while the RT loop and the sheet formed by the N- and C-terminal strands are very rarely populated, suggesting that they are the last elements to be structured in the protein (the contact map for the native protein is shown above the diagonal for comparison).

Table 2 Kinetic parameters of src SH3 folding[1]

	Mutant	ln(k_f)	ln(k_u)	m_f	m_u	$\Delta\Delta G_u$	Φ_f
	WT	3.55 ± 0.03	1.07 ± 0.04	1.02 ± 0.019	0.54 ± 0.013	NA	NA
	T9A	3.67 ± 0.03	2.35 ± 0.04	1.01 ± 0.037	0.46 ± 0.017	-0.64 ± 0.08	-0.11 ± 0.04
β1	F10A	3.41 ± 0.03	2.39 ± 0.03	1.06 ± 0.029	0.51 ± 0.012	-0.84 ± 0.07	0.10 ± 0.03
	F10I	3.67 ± 0.04	4.06 ± 0.03	1.08 ± 0.11	0.44 ± 0.091	-1.65 ± 0.17	-0.05 ± 0.02
	V11A	3.45 ± 0.04	3.79 ± 0.03	0.85 ± 0.057	0.55 ± 0.033	-1.64 ± 0.12	0.03 ± 0.02
	A12G	3.46 ± 0.03	2.73 ± 0.05	1.08 ± 0.049	0.47 ± 0.026	-1.00 ± 0.09	0.05 ± 0.02
	L13A	3.62 ± 0.02	3.87 ± 0.06	1.20 ± 0.098	0.36 ± 0.02	-1.49 ± 0.13	-0.03 ± 0.01
	Y14A	3.59 ± 0.03	1.68 ± 0.03	0.96 ± 0.042	0.50 ± 0.011	-0.31 ± 0.06	-0.08 ± 0.09
	D15A	3.70 ± 0.03	2.02 ± 0.08	0.98 ± 0.051	0.41 ± 0.03	-0.43 ± 0.13	-0.22 ± 0.10
	Y16A	3.40 ± 0.07	4.94 ± 0.06	–[2]	0.39 ± 0.018	-2.27 ± 0.26	0.03 ± 0.03
RT	Y16F	3.54 ± 0.04	1.37 ± 0.02	0.97 ± 0.036	0.65 ± 0.093	-0.18 ± 0.10	–[3]
loop	S18A	3.80 ± 0.04	0.47 ± 0.06	0.96 ± 0.028	0.55 ± 0.051	0.52 ± 0.10	0.28 ± 0.06
	R19A	3.64 ± 0.04	1.22 ± 0.07	1.08 ± 0.034	0.61 ± 0.017	-0.07 ± 0.08	–[3]
	T20A	3.69 ± 0.03	1.10 ± 0.04	1.01 ± 0.033	0.52 ± 0.017	0.06 ± 0.07	–[3]
	T22A	3.60 ± 0.03	1.14 ± 0.03	1.00 ± 0.03	0.52 ± 0.016	-0.01 ± 0.07	–[3]
	D23A	3.43 ± 0.05	1.91 ± 0.07	1.06 ± 0.083	0.51 ± 0.043	-0.56 ± 0.13	0.13 ± 0.07
	L24A	2.76 ± 0.03	3.44 ± 0.03	1.45 ± 0.06	0.45 ± 0.0089	-1.79 ± 0.09	0.26 ± 0.01
	S25A	3.49 ± 0.04	2.40 ± 0.03	0.95 ± 0.041	0.58 ± 0.04	-0.82 ± 0.08	0.03 ± 0.04
	F26A	2.17 ± 0.03	3.21 ± 0.04	–[2]	0.40 ± 0.0096	-1.97 ± 0.10	0.40 ± 0.01
	K27A	3.60 ± 0.04	1.79 ± 0.05	1.01 ± 0.051	0.65 ± 0.045	-0.44 ± 0.11	-0.06 ± 0.09
	K28A	3.73 ± 0.03	1.45 ± 0.04	0.88 ± 0.017	0.49 ± 0.022	-0.09 ± 0.07	–[3]
diverging	G29A	2.30 ± 0.04	2.74 ± 0.03	1.54 ± 0.088	0.43 ± 0.017	-1.66 ± 0.12	0.44 ± 0.02
turn	E30A	1.50 ± 0.03	2.34 ± 0.03	1.09 ± 0.054	0.65 ± 0.027	-1.94 ± 0.13	0.62 ± 0.02
	R31A	3.43 ± 0.03	1.45 ± 0.10	1.00 ± 0.06	0.57 ± 0.035	-0.32 ± 0.12	0.23 ± 0.08
β2	L32A	1.40 ± 0.08	2.61 ± 0.11	–[2]	–[2]	-2.26 ± 0.37	0.55 ± 0.05
	L32V	3.10 ± 0.04	2.68 ± 0.03	1.03 ± 0.05	0.56 ± 0.031	-1.21 ± 0.11	0.22 ± 0.02
	Q33A	3.17 ± 0.04	1.00 ± 0.04	1.11 ± 0.033	0.55 ± 0.027	-0.21 ± 0.09	–[4]
	I34A	1.40 ± 0.04	-0.63 ± 0.03	1.39 ± 0.027	0.51 ± 0.03	-0.32 ± 0.12	–[4]
	I34V	3.09 ± 0.05	0.75 ± 0.09	1.16 ± 0.093	0.50 ± 0.019	-0.09 ± 0.12	–[4]
n-src	V35A	2.6 ± 0.06	1.33 ± 0.05	1.27 ± 0.046	0.59 ± 0.041	-0.77 ± 0.12	0.77 ± 0.05
loop	N36A	3.64 ± 0.05	1.50 ± 0.05	1.12 ± 0.035	0.47 ± 0.017	-0.20 ± 0.09	–[3]
	N37A	3.91 ± 0.03	1.30 ± 0.04	0.94 ± 0.032	0.59 ± 0.023	0.07 ± 0.06	–[4]
	G40A	2.99 ± 0.05	1.02 ± 0.06	0.92 ± 0.037	0.55 ± 0.036	-0.28 ± 0.08	–[4]
	W42A	3.03 ± 0.05	2.83 ± 0.02	1.62 ± 0.048	0.45 ± 0.0084	-1.29 ± 0.10	0.25 ± 0.03
	W43A	3.24 ± 0.03	2.97 ± 0.06	1.16 ± 0.069	0.35 ± 0.025	-1.20 ± 0.11	0.15 ± 0.02
	W43I	4.45 ± 0.05	3.30 ± 0.03	1.10 ± 0.081	0.62 ± 0.073	-0.77 ± 0.13	–[4]
β3	L44A	2.10 ± 0.05	2.52 ± 0.05	1.88 ± 0.065	0.41 ± 0.0095	-1.64 ± 0.15	0.54 ± 0.03
	A45G	1.71 ± 0.07	0.86 ± 0.06	1.70 ± 0.044	0.39 ± 0.011	-0.92 ± 0.15	1.20 ± 0.08
	H46A	3.47 ± 0.03	2.00 ± 0.02	0.99 ± 0.029	0.58 ± 0.016	-0.62 ± 0.06	0.08 ± 0.04
	S47A	1.23 ± 0.04	1.29 ± 0.03	1.50 ± 0.035	0.44 ± 0.0074	-1.46 ± 0.09	0.95 ± 0.03
	L48A	2.82 ± 0.04	1.38 ± 0.03	1.20 ± 0.059	0.51 ± 0.019	-0.61 ± 0.08	0.72 ± 0.04
distal	S49A	2.98 ± 0.06	0.07 ± 0.06	1.20 ± 0.061	0.61 ± 0.026	0.18 ± 0.11	–[4]
hairpin	T50A	0.99 ± 0.05	1.59 ± 0.02	1.84 ± 0.054	0.47 ± 0.012	-1.79 ± 0.10	0.86 ± 0.02
	G51A	1.39 ± 0.04	1.02 ± 0.06	1.51 ± 0.11	0.41 ± 0.019	-1.21 ± 0.14	1.06 ± 0.06
	Q52A	3.29 ± 0.03	1.41 ± 0.08	1.03 ± 0.059	0.49 ± 0.021	-0.35 ± 0.12	0.45 ± 0.09
	T53A	2.33 ± 0.06	1.65 ± 0.04	1.38 ± 0.048	0.56 ± 0.021	-1.11 ± 0.11	0.68 ± 0.03
β4	G54A	3.79 ± 0.02	4.59 ± 0.05	1.60 ± 0.17	0.36 ± 0.02	-1.81 ± 0.12	-0.08 ± 0.01
	Y55A	2.10 ± 0.04	2.34 ± 0.04	1.64 ± 0.063	0.39 ± 0.0062	-1.52 ± 0.10	0.56 ± 0.02
	I56A	1.34 ± 0.03	2.02 ± 0.02	1.64 ± 0.067	0.46 ± 0.016	-1.84 ± 0.10	0.71 ± 0.02
	P57A	2.98 ± 0.04	2.89 ± 0.04	1.36 ± 0.098	0.45 ± 0.014	-1.36 ± 0.11	0.24 ± 0.02
3_{10}	S58A	4.06 ± 0.03	1.14 ± 0.05	0.99 ± 0.034	0.58 ± 0.023	0.24 ± 0.08	–[4]
helix	N59A	3.55 ± 0.03	1.28 ± 0.04	0.87 ± 0.041	0.58 ± 0.016	-0.14 ± 0.07	–[3]
	Y60A	3.48 ± 0.05	0.65 ± 0.04	1.06 ± 0.034	0.47 ± 0.034	0.23 ± 0.09	–[3]
	V61A	3.67 ± 0.04	3.29 ± 0.02	1.15 ± 0.044	0.44 ± 0.013	-1.18 ± 0.09	-0.06 ± 0.03
β5	A62G	3.59 ± 0.04	1.97 ± 0.03	1.12 ± 0.045	0.55 ± 0.019	-0.53 ± 0.08	-0.02 ± 0.07
	P63A	3.64 ± 0.04	1.44 ± 0.08	1.04 ± 0.062	0.48± 0.031	-0.14 ± 0.11	–[3]
	S64A	3.66 ± 0.03	0.40 ± 0.04	0.95 ± 0.026	0.59 ± 0.018	0.44 ± 0.06	0.14 ± 0.05
	D65A	3.69 ± 0.03	0.98 ± 0.05	0.92 ± 0.032	0.57 ± 0.02	0.13 ± 0.07	–[3]

[1]All experiments were done at pH 6 and 295 K. Kinetic measurements were done by stopped-flow fluorescence. Rate of folding (k_f) is reported at 0.3 M Gnd; rate of unfolding (k_u) is reported at 4 M Gnd to avoid extrapolation. $\Delta\Delta G_u$, Φ_f and standard errors were calculated as described in the Methods section.

[2]Parameters could not be reliably measured.

[3]Mutation has no (or very small) effect on stability, that is, $\Delta\Delta G_u \leq 0.20$ kcal mol^{-1}.

[4]Mutation either increases or decreases both k_f and k_u.

High-temperature MD unfolding simulations have provided insights into the folding of a number of small proteins[27–30]. Tsai *et al.*[22] carried out 30 independent simulations of src SH3 domain unfolding, and analyzed the order in which the structural elements are disrupted in the unfolding process. Overall, the hierarchy of unfolding was consistent with, but less pronounced than, the hierarchy observed in the experiments and in the *ab initio* simulations: the interactions between the N- and C-terminal strands were lost earlier than those within the distal loop β-hairpin, the n-src loop, and between the diverging turn and the distal loop β-hairpin.

Although the overall features of the simulations were consistent with the experimental results, there also were some inconsistencies. While residues in the three-stranded sheet made extensive contacts in the *ab initio* simulations and have high Φ_f values in the experiments, the C-terminus also made numerous contacts in the simulations (Fig. 3*a*), but contains mainly low Φ_f values. In the MD simulations, the RT loop remained ordered until quite late in the unfolding process. These discrepancies notwithstanding, the overall concordance between the simulations and the experiments is quite intriguing given that the MD simulations were carried out at 498 K and the *ab initio* simulations do not explicitly model side chains, and sugests that the hierarchy to folding of the SH3 domain is determined by fairly coarse-grained features of the structure.

Native state topology-based model calculations

To isolate those features of the native topology responsible for determining folding mechanisms, we recently developed a simple native state topology-based model of the folding free energy landscape and folding process[1,31]. In this model, the folding landscape is approximated by considering only conformations in which each residue is either ordered as in the native structure or completely disordered, and all ordered residues occur in one or two contiguous stretches of the protein sequence. The free energy of each of these conformations is determined by the balance between attractive native interactions, taken to be proportional to the surface area buried within the ordered region in the native structure, and the entropic cost of chain ordering, a function of the number of residues ordered and the loop length between the ordered segments.

We use this simple approach to model the folding free energy landscape of the three SH3 domains whose folding mechanisms have been probed by mutation: the src SH3 domain, the fyn SH3 domain (A. Davidson, pers. comm.), and the α-spectrin SH3 domain[6,12]. As a control, the same calculations were carried out on the α-spectrin SH3 permutant found by Serrano and coworkers[12] to have a significantly changed folding transition state. A natural reaction coordinate in this model is simply the fraction of residues ordered as in the native state, N_f. To determine the order in which the different parts of the protein fold as N_f increases, we enumerated all configurations allowed by the model with a particular value of N_f, determined their free energies, and computed the Boltzmann weighted frequency of ordering each residue[31] (Fig. 3*b*). Close to the unfolded state ($N_f = 0$) most residues have low frequency of ordering (black color), while close to the native state ($N_f = 1$) almost all residues are ordered (white color). There are interesting similarities in the hierarchies of structure formation in the three native SH3 domains obtained with this model (Fig. 3*b*). The first regions of the proteins to become ordered are the three hairpin loops (the distal loop β-hairpin, the RT loop and the n-src loop). By $N_f \sim 0.5$, the predominant region ordered in all three proteins is the three-stranded sheet formed by the distal loop β-hairpin and the n-src loop. The decrease in the relative population of the RT loop occurs because ordering additional residues increases the entropic cost of structure formation without significant increases in the attractive native interactions; in contrast, the ordering of the residues in the three-stranded sheet formed by the n-src loop and the distal loop β-hairpin produces significant gains in attractive interactions (the three-stranded sheet has a much higher density of stabilizing interactions than other portions of the protein of similar length). For all three proteins, the first and last strands become ordered very late in the folding process, consistent with the fact that they are stabilized primarily by nonlocal interactions. The similarities of the plots are consequences of the similarities in the topology of the three proteins. Notably, the hierarchy of folding is significantly altered in the SH3 domain circular permutant (Fig. 3*b*).

Overall, the hierarchy of structure formation observed in this simple model is consistent with the experimental results (Fig. 1): the residues with high Φ_f values in the src SH3 domain lie in the central three-stranded sheet, and for α-spectrin[6] and fyn (A. Davidson, pers. comm.), mutations in the distal loop β-hairpin have high Φ_f values. The lack of treatment of local sequence structure propensities may account for the roughly equal tendencies of the n-src loop and the distal loop β-hairpin to form in the model; the higher Φ_f values in the distal loop β-hairpin in src and α-spectrin[6] may reflect more complete ordering of this structure in the transition state due to stabilizing local interactions such as hydrogen bonds not included in the model. It should be emphasized that the two-segment model does not simply identify the longest contiguous stretch of interacting residues; for example, in barnase the N-terminal helix is correctly predicted to associate with the C-terminal sheet in the transition state[31].

In summary, the similarity in the hierarchy of folding observed experimentally in the src and α-spectrin SH3 domains, in the *ab initio* and MD simulations, and in the simple model calculations suggests that the folding mechanism of SH3 domains is largely determined by the topology of the native protein. The success of the simple model in reproducing the hierarchy observed both experimentally and in the simulations suggests that the folding process of this protein is largely determined by the balance between the entropic cost of chain ordering and the formation of attractive native interactions; nonnative interactions and conformations (that is, kinetic traps) appear to play a relatively minor role in shaping the folding process. The structural polarization of the SH3 domain folding transition state can be viewed as a consequence of the low free energy cost of ordering the low contact order[7] central three-stranded sheet, relative to the much higher contact order sheet formed by the N- and C-termini together with the RT loop. The importance of the computational work described in this paper in supporting this hypothesis may be seen by considering the alternative hypothesis that structural polarization in the transition state ensemble is a consequence of inhomogeneities in inter-residue interaction strengths: the strongest interactions are the last to break during unfolding and the most likely to nucleate the refolding process. Since the distal loop β-hairpin has the most extensive intraloop hydrogen bonding, if only the experimental data were available it could equally well be argued that the origin of structural polarization of the SH3 transition state was the greater stabilization of the distal loop β-hairpin relative to the

other structural elements in the protein. The *ab initio* folding simulations, however, have no prior knowledge that the interactions within the distal loop β-hairpin are stronger than in the other loops, and the simple free energy landscape model does not consider hydrogen bonding at all. Thus, the fact that a similar hierarchy to structure formation is observed in the calculations and experiments helps to distinguish between two hypotheses that are equally consistent with the experimental data.

The accompanying papers from the Dobson[32] and Serrano[33] groups strongly support the idea that native state topology is a dominant determinant of protein folding mechanisms. Martinez and Serrano[33] show that the folding transition state of the α-spectrin SH3 domain is similar to that of the src SH3 domain and is not significantly altered by changes in pH that produce large changes in stability. Chiti *et al.*[32] show that folding transition state structure is conserved in a second pair of proteins with similar native structures but with only 13% sequence identity: acylphosphatase and the activation domain of procarboxypeptidase 2. Chiti *et al.*[32] also show that the correlation between folding rates and contact order observed among two-state folding proteins generally also holds within a set of five nonhomologous proteins that exhibit the AcP topology.

The combination of experiment, simulation and theory employed in this paper, together with comparisons of the folding of structurally related proteins such as those in the accompanying papers, has the potential to distinguish the robust features of the folding process from those dependent on high-resolution detail, and to trace the origins of these robust features to basic physical principles. We believe that this integration of complementary approaches will be critical for obtaining a complete understanding of the folding process.

Methods

Mutagenesis. Mutagenesis was accomplished using the Quick Change site-directed mutagenesis kit (Stratagene). Plasmids harboring the point mutations were transformed into BL21 cells, and protein was overexpressed and purified[5]. The His tag was not removed for the purposes of this study. All mutants were sequenced to ensure that the mutagenesis was successful and the purified proteins were analyzed by mass spectrometry to confirm that each mutation was the expected one.

Biophysical analysis. In all experiments, protein solutions were made in 50 mM sodium phosphate, pH 6, and the temperature was held constant at 295 K. The stability of the point mutants was assessed by guanidine (Gnd) denaturation using either circular dichroism (CD) or fluorescence as reported[5]. The kinetics of folding and unfolding were followed by fluorescence on a Bio-Logic SFM-4 stopped-flow instrument. The unfolding reaction for the wild type protein was determined to behave as a two-state process[8], and the kinetic and equilibrium data for the mutants were fit to a two-state model. Equilibrium data (not shown) were generally in agreement with the kinetic estimates of stability for the less destabilized mutants.

Φ value analysis. Φ_f values were calculated only for mutants that were destabilized by more than 0.2 kcal mol^{-1} relative to the wild type protein. In order to avoid extrapolations, we compared folding rates at 0.3 M Gnd and unfolding rates at 4 M Gnd. In calculating $\Delta\Delta G$ for each mutant we assumed that it is independent of the denaturant concentration, which is warranted since the m values for the mutants are not very different from wild type. $\Delta\Delta G$ and Φ_f were computed using $\Delta\Delta G = -RT(\ln(k_{0.3M}^{wt}/k_{0.3M}^{mut}) + \ln(k_{4M}^{mut}/k_{4M}^{wt}))$ and $\Phi_f = -RT\ln(k_{0.3M}^{wt}/k_{0.3M}^{mut})/\Delta\Delta G$. For error analysis, we decided on a procedure that makes use of the many independent measurements in the linear portions of the V curves shown in Fig. 2. The estimates and confidence regions for $\ln(k_{0.3M}^{wt}/k_{0.3M}^{mut})$ and $\ln(k_{4M}^{mut}/k_{4M}^{wt})$ were obtained by simultaneously fitting the linear portions of the mutant and wild type V curves to $\ln k_f(Gnd)^{wt} = \ln k_f(Gnd)^{mut} + \delta_f$ and $\ln k_u(Gnd)^{wt} = \ln k_u(GND)^{mut} + \delta_u$. The error estimates for the Φ_f values presented in Table 2 represent 95% confidence intervals (roughly twice the standard deviation) for the Φ_f value, generated by repeatedly (10,000 times) sampling from the δ_f and δ_u distributions and recomputing the Φ_f values using $\Phi_f = \delta_f / (\delta_f + \delta_u)$.

***Ab initio* folding simulations.** The *ab initio* folding method, ROSETTA, utilizes a backbone plus side chain centroid-based representation of the chain; local interactions are satisfied by building structures up from short (three- and nine-residue) segments of known structures with sequences similar to those of the sequence being folded[24], while the nonlocal interactions that stabilize proteins are treated using a low-resolution scoring function with terms representing hydrophobic burial, strand pairing and specific pair interactions such as charge pairing and disulfide bonding[25]. A Monte Carlo simulated annealing strategy is used to sample conformational space; a move consists of a substitution of a three- or nine-residue segment of the chain by a randomly chosen fragment from a known structure with a similar local sequence. The protocol used to simulate the SH3 domain folding here was the same as that used in our CASP3 structure predictions[23]. All SH3 domain structures were removed from the data base of short fragments used for building up conformations. Because of the very large size of the conformational space, the trajectories and final structures for different runs can vary considerably. A total of 500 independent simulations were carried out, and all conformations from the 20 trajectories that produced structures within 4.5 Å r.m.s.d. of the native structure were combined. The frequency of each contact (defined as a pair of side chain centroids within 8 Å) in the pooled set of conformations was then computed.

Simple model calculations. The folding free energy landscape of the SH3 domain was approximated using the two-segment model described[31]. The free energy landscapes of the src, spectrin and fyn SH3 domains, and the 47–48 circular permutant of the spectrin SH3 domain were approximated by considering only configurations in which (i) each residue is fully ordered as in the native state or fully disordered, and (ii) the ordered residues occur in one or two contiguous stretches of the sequence. The free energy of each configuration was computed from the equation $F = -\gamma\Delta ASA + \alpha RTN + \beta RT\ln(\Delta L)$; all parameters were taken from the literature or from simple off-lattice calculations. In the first term, which represents the favorable interactions made in the partially ordered configuration, ΔASA is the difference in exposed surface area between the partially ordered configuration and the unfolded state (estimated from the sum of the native tripeptide surface areas) and $\gamma = 16$ cal mol^{-1} Å^{-2}. In the second term, which represents the entropic cost of ordering each residue in the ordered segments, N is the number of residues ordered and $\alpha = 2.9$. In the third term, which represents the entropic cost of closing the loop between the two ordered segments[34], ΔL is the length of the loop and $\beta = 1.8$.

Acknowledgments

We thank members of the Baker group for useful comments on the manuscript, and L. Serrano, C. Dobson and their coworkers for sharing their manuscripts before publication. This work was supported by grants from the NIH and the ONR and Young Investigator awards to D.B. from the NSF and the Packard Foundation and by the Molecular Biophysics Training Grant from the NIH to E.A.

Correspondence should be addressed to D.B.
email: dabaker@u.washington.edu

Received 30 April, 1999; accepted 1 September, 1999.

1. Alm, E. & Baker, D. *Curr. Opin. Struct. Biol.* **9**, 189–196 (1999).
2. Riddle, D.S. *et al. Nature Struct. Biol.* **4**, 805–809 (1997).
3. Kim, D.E., Gu, H. & Baker, D. *Proc. Natl. Acad. Sci USA* **95**, 4982–4986 (1998).
4. Perl, D. *et al. Nature Struct. Biol.* **5**, 229–235 (1998).
5. Grantcharova, V.P., Riddle, D.S., Santiago, J.V. & Baker, D. *Nature Struct. Biol.* **5**, 714–720 (1998).
6. Martinez, J.C., Pisabarro, M.T. & Serrano, L. *Nature Struct. Biol.* **5**, 721–726 (1998).
7. Plaxco, K.W., Simons, K. & Baker, D. *J. Mol. Biol.* **277**, 985–994 (1998).
8. Grantcharova, V.P. & Baker, D. *Biochemistry* **36**, 15685–15692 (1998).
9. Plaxco, K.W. *et al. Biochemistry* **37**, 2529–2537 (1998).
10. Guijarro, J.I., Morton, C., Plaxco, K.W., Campbell, I.D. & Dobson, C.M. *J. Mol. Biol.* **276**, 657–667 (1998).
11. Viguera, A.R., Martinez, J.C., Filimonov, V.V., Mateo, P.L. & Serrano, L. *Biochemistry* **33**, 2142–2150 (1994).
12. Viguera, A.R., Serrano, L. & Wilmanns, M. *Nature Struct. Biol.* **3**, 874–879 (1996).
13. Maxwell, K.L. & Davidson, A.R. *Biochemistry* **37**, 16172–16182 (1998).
14. Lim, W.A., Fox, R.O. & Richards, F.M. *Protein Sci.* **3**, 1261–1266 (1994).
15. Fersht, A.R. *Curr. Opin. Struct. Biol.* **5**, 79–84 (1994).
16. Itzhaki, L.S., Otzen, D.E. & Fersht, A.R. *J. Mol. Biol.* **254**, 260–288 (1995).
17. Milla, M.E., Brown, B.M., Waldburger, C.D. & Sauer, R.T. *Biochemistry* **34**, 13914–13919 (1995).
18. Serrano, L., Matouschek, A. & Fersht, A.R. *J. Mol. Biol.* **224**, 805–818 (1992).
19. Lopez-Hernandez, E. & Serrano, L. *Folding & Design* **1**, 43–55 (1995).
20. Kragelund, B.B. *et al. Nature Struct. Biol.* **6**, 594–601 (1999).
21. Yi, Q., Bystroff, C., Rajagopal, P., Klevit, R.E. & Baker, D. *J. Mol. Biol.* **283**, 293–300 (1998).
22. Tsai, J., Levitt, M. & Baker, D. *J. Mol. Biol.* **291**, 215–225 (1999).
23. Simons, K.T., Bonneau, R., Ruczinski, I. & Baker, D. *Proteins Struct. Funct. Genet.*, **in the press** (1999).
24. Simons, K.T., Kooperberg, C., Huang, E. & Baker, D. *J. Mol. Biol.* **268**, 209–25 (1997).
25. Simons, K.T. et al. Proteins Struct. Funct. Genet. **34**, 82–95 (1999).
26. Du, R., Pande, V.S., Grosberg, A.Y., Tanaka, T. & Shaknovich, E.I. *J. Chem. Phys.* **108**, 334–350 (1998).
27. Li, A. & Daggett, V. *J. Mol. Biol.* **257**, 412–429 (1996).
28. Bond, C.J., Wong, K.B., Clarke, J. Fersht, A.R. & Daggett, V. *Proc. Natl. Acad. Sci. USA* **94**, 13409–13413 (1997).
29. Alonso, D.O. & Daggett, V. *Protein Sci.* **7**, 860–874 (1998).
30. Lazaridis, T. & Karplus, M. *Science* **278**, 1928–1931 (1997).
31. Alm, E. & Baker, D. *Proc. Natl. Acad. Sci. USA*, **96**, 11305–11310 (1999).
32. Chiti, F. *et al.* Nature Struct. Biol. **6**, 1005–1009 (1999).
33. Martinez, J.C. & Serrano, L. *Nature Struct. Biol.* **6**, 1010–1016 (1999).
34. Jacobson, H. & Stockmayer, W.H. *J. Chem. Phys.* **18**, 1600–1606 (1950).
35. Xu, W., Harrison, S.C., & Eck, M.J. *Nature* **385**, 595–602 (1997).
36. Kraulis, P.J. *J. Appl. Crystallogr.* **24**, 946–950 (1991).

Long-range order in the src SH3 folding transition state

Viara P. Grantcharova, David S. Riddle*, and David Baker†

Department of Biochemistry, University of Washington, Seattle, WA 98195

Communicated by Robert L. Baldwin, Stanford University Medical Center, Stanford, CA, April 19, 2000 (received for review February 21, 2000)

One of the outstanding questions in protein folding concerns the degree of heterogeneity in the folding transition state ensemble: does a protein fold via a large multitude of diverse "pathways," or are the elements of native structure assembled in a well defined order? Herein, we build on previous point mutagenesis studies of the src SH3 by directly investigating the association of structural elements and the loss of backbone conformational entropy during folding. Double-mutant analysis of polar residues in the distal β-hairpin and the diverging turn indicates that the hydrogen bond network between these elements is largely formed in the folding transition state. A 10-glycine insertion in the n-src loop (which connects the distal hairpin and the diverging turn) and a disulfide crosslink at the base of the distal β-hairpin exclusively affect the folding rate, showing that these structural elements are nearly as ordered in the folding transition state as in the native state. In contrast, crosslinking the base of the RT loop or the N and C termini dramatically slows down the unfolding rate, suggesting that dissociation of the termini and opening of the RT loop precede the rate-limiting step in unfolding. Taken together, these results suggest that essentially all conformations in the folding transition state ensemble have the central three-stranded β-sheet formed, indicating that, for the src homology 3 domain, there is a discrete order to structure assembly during folding.

One of the major differences between the "old" and the "new" views of protein folding is the degree of heterogeneity in the folding transition state ensemble (1). In the limit of a single folding pathway, all folding trajectories are presumed to undergo similar conformational transitions, whereas in the limit of a perfectly symmetric folding "funnel," a vast number of different trajectories lead to the native state. Studies of simple lattice models have suggested both "single" and "multiple" folding nuclei scenarios (2, 3). For small proteins that fold without detectable intermediates, characterization of the folding transition state by mutational analysis is perhaps the best available approach to addressing this issue. This method, pioneered by Fersht and coworkers (4, 5), has proven extremely powerful in providing site-specific information about structure at the rate-limiting step for folding (6–11). It probes the formation of side chain–side chain interactions in the transition state by deleting parts of individual residues and assessing the effect on folding kinetics. There are, however, two shortcomings of this approach: (*i*) residues are often involved in multiple interactions, and point mutagenesis does not distinguish which of these are important in the transition state; and (*ii*) the conformation of the peptide backbone can be deduced only indirectly. To go beyond these limitations, in the present study, we employ double-mutant analysis to probe side chain–side chain interactions between structural elements and a 10-glycine insertion and disulfide crosslinks to test backbone ordering and the association of entire structural elements at the transition state.

Previously, we studied the structure of the transition state for folding of the 57-residue src SH3 domain by characterizing the kinetic consequences of a large number of point mutations and found that the rate-limiting step in folding involves formation of the distal β-hairpin and the diverging turn (Fig. 1*A*; refs. 12 and 13). In this study, we investigate long-range order in the transition state by: (*i*) double-mutant analysis of a hydrogen bond network between the distal β-hairpin and the diverging turn to probe their association in the transition state; (*ii*) a 10-glycine insertion in the n-src loop to investigate its conformational rigidity and thus the association of the distal β-hairpin and the diverging turn, which the n-src loop connects; (*iii*) disulfide crosslinking the distal β-hairpin and the RT loop to probe the extent of closure of the two hairpin loops; and (*iv*) disulfide crosslinking of the N and C termini to probe the association of the terminal strands in the transition state.

Methods

Mutagenesis. Point mutagenesis was accomplished with the Quick Change Site-Directed mutagenesis kit (Stratagene). The glycine insertion mutant was constructed by using PCR cassette mutagenesis with primers coding for the 10 glycines. Plasmids harboring the mutations were transformed into BL21 cells, and protein was overexpressed and purified (12). The His tag was not removed for the purposes of this study. All mutants were sequenced to ensure that the mutagenesis was successful, and the purified proteins were analyzed by mass spectrometry to confirm protein identity.

Disulfide Crosslinking. Residues mutated to cysteine were chosen to satisfy the geometric requirements for disulfide bond formation: Cα–Cα (4.4–6.8 Å), Cβ–Cβ (3.5–4.5 Å), and dihedral angle close to 90° (14). The residues chosen were previously determined to have a very small or no effect on the rate of folding and stability (13). W43C is the only mutation that is likely to affect stability significantly (as judged from the W43A mutation, $\Delta\Delta G$ = 1.2 kcal/mol). Disulfide bonds were oxidized in the presence of 20 mM $K_3Fe(CN)_6$ for 10 min at room temperature. Reactions were performed in the dark because $K_3Fe(CN)_6$ is light sensitive. Disulfide formation was confirmed with Ellman's reagent.

Biophysical Analysis. Protein solutions (100 μM) were made in 50 mM sodium phosphate (pH 6), and the temperature was held constant at 295 K. For wild type (WT) and the S47A mutant, experiments were also performed in 50 mM NaPi (pH 3) at 295 K. To reduce the disulfide-crosslink mutants, they were incubated in 10 mM DTT for 1 h, and the same concentration of reducing agent was present throughout the kinetic experiments. The kinetics of folding and unfolding were followed by tryptophan fluorescence on a Bio-Logic SFM-4 stopped-flow instrument (Molecular Kinetics, Pullman, WA). The unfolding reaction for the WT protein can be modeled as a two-state process (15), and the kinetic and equilibrium data for the mutants were

Abbreviations: SH3, src homology 3; WT, wild type; Gnd, guanidine.

*Present address: Department of Immunology, Mayo Clinic, Rochester, MN 55904

†To whom reprint requests should be addressed. E-mail: dabaker@u.washington.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

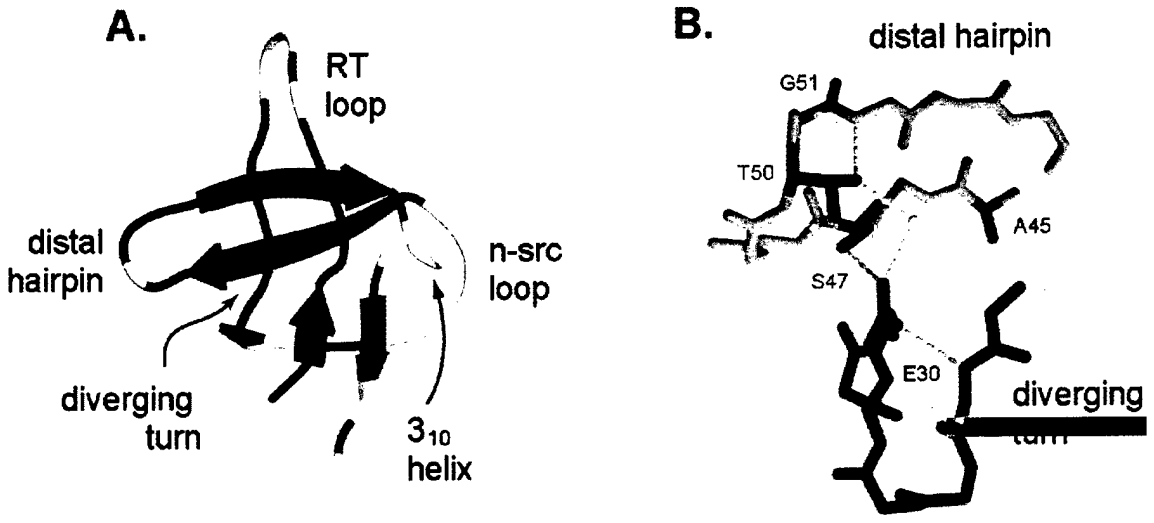


Fig. 1. (*A*) Structure of the src SH3 domain (1fmk.pdb) colored by previously reported Φ_F-values (13) on a continuous scale from red ($\Phi_F = 1$) to blue ($\Phi_F = 0$). Residues at which mutations increase or decrease both k_f and k_u are colored in yellow. The graphic was generated with MOLSCRIPT (33) and RASTER3D (34, 35). (*B*) Structure of the hydrogen bond network between the β-distal hairpin and the diverging turn (MIDAS; refs. 36 and 37). Residues included in the double-mutant cycles are shown in red.

fit to a two-state model. For each mutant, the free energy of folding is calculated as:

$$\Delta G_{U\text{-}F} = RT\ln(k_f/k_u), \quad [1]$$

where k_f and k_u are the rates of folding and unfolding, respectively, in the absence of denaturant. The differences in the free energy of folding ($\Delta\Delta G_{U\text{-}F}$) and in the folding activation energy ($\Delta\Delta G_{U\text{-}\ddagger}$) between the WT protein and each mutant are calculated as:

$$\Delta\Delta G_{U\text{-}F} = RT[\ln(k_f^{wt}/k_f^{mut}) + \ln(k_u^{mut}/k_u^{wt})] \quad \text{and}$$

$$\Delta\Delta G_{U\text{-}\ddagger} = RT\ln(k_f^{wt}/k_f^{mut}), \quad [2]$$

where k_f and k_u are the rates of folding and unfolding, respectively, at denaturant concentrations experimentally accessible for that mutant. The parameter Φ_F is defined as

$$\Phi_F = \Delta\Delta G_{U\text{-}\ddagger}/\Delta\Delta G_{U\text{-}F} \quad [3]$$

and is interpreted as the fraction of the mutated residue's interactions that are formed in the transition state. A Φ_F-value of 1 indicates that all of a residue's interactions are formed in the transition state, whereas a Φ_F of 0 means that the residue does not make stabilizing interactions in the transition state (5). In the case of the double mutants, a Φ_F-value for the pairwise interaction can be determined similarly:

$$\Phi_F^{int} = \Delta\Delta G_{U\text{-}\ddagger}^{int}/\Delta\Delta G_{U\text{-}F}^{int}. \quad [4]$$

Loop Entropy Estimates. The change in the free energy of the unfolded state as a result of loop insertion or disulfide crosslinking can be estimated from polymer theory (16):

$$\Delta G = -RT\ln(L/L_o), \quad [5]$$

where L_o and L are loop lengths before and after the modification, respectively. In the case of the 10-glycine insertion in the n-src loop, the original loop length is 5 and after the insertion it is 15. In the case of the three disulfide crosslinks, which generate a loop in the protein that did not exist previously, L_o is taken to be 1, and L is the length of the loop enclosed by the crosslink.

Results

Hydrogen Bond Network Between the Distal β-Hairpin and the Diverging Turn. Mutagenic analysis of the SH3 domain folding transition state revealed the clustering of structured residues in the distal β-hairpin and the diverging turn (12). Mutagenesis also suggested that these elements might interact with each other at the rate-limiting step, because mutations in the hydrogen bond network between the distal β-hairpin and the diverging turn (Fig. 1*B*) had a dramatic effect on the rate of folding. In this study, we performed double-mutant cycles on these hydrogen bond network residues (S47 and T50 in the distal β-hairpin and E30 in the diverging turn) to quantify interaction energies at the transition state (4). Both E30A_S47A (Fig. 2*A*) and E30A_T50A (Fig. 2*B*) double mutants are considerably less destabilized than expected from the sum of the single-mutant effects (Table 1): in the native state, the interaction energy between the two mutated residues is 1.02 kcal/mol for E30A and S47A (Fig. 2*D*) and 1.08 kcal/mol for E30A and T50A. A large fraction of this interaction energy is present at the folding transition state: 0.78 kcal/mol for E30 and S47 and 0.83 kcal/mol for E30 and T50, yielding interaction Φ_F-values of 0.76 and 0.77, respectively. The finding that these hydrogen bonds are already formed at the transition state confirms directly the association of the distal β-hairpin and the diverging turn in the transition state inferred from the point mutagenesis experiments.

The E30A mutation removes a large portion of this buried residue and is therefore expected to be quite disruptive. As a less drastic probe of the interactions in the hydrogen bond network, we compare the kinetic effect of the S47A mutation at pH 3 and pH 6 (Fig. 2*C* and Table 1). At pH 3, the carboxyl group of E30 probably is partially protonated (depending on the local pKa), thus disrupting some of its interactions with the distal β-hairpin (S47). The effect of S47A on stability and the rate of folding is smaller at pH 3 than at pH 6, but the Φ_F-value is still 1. This effect is consistent with the idea that some of the interactions between S47 and E30 present at the transition state at pH 6 can be disrupted by low pH.

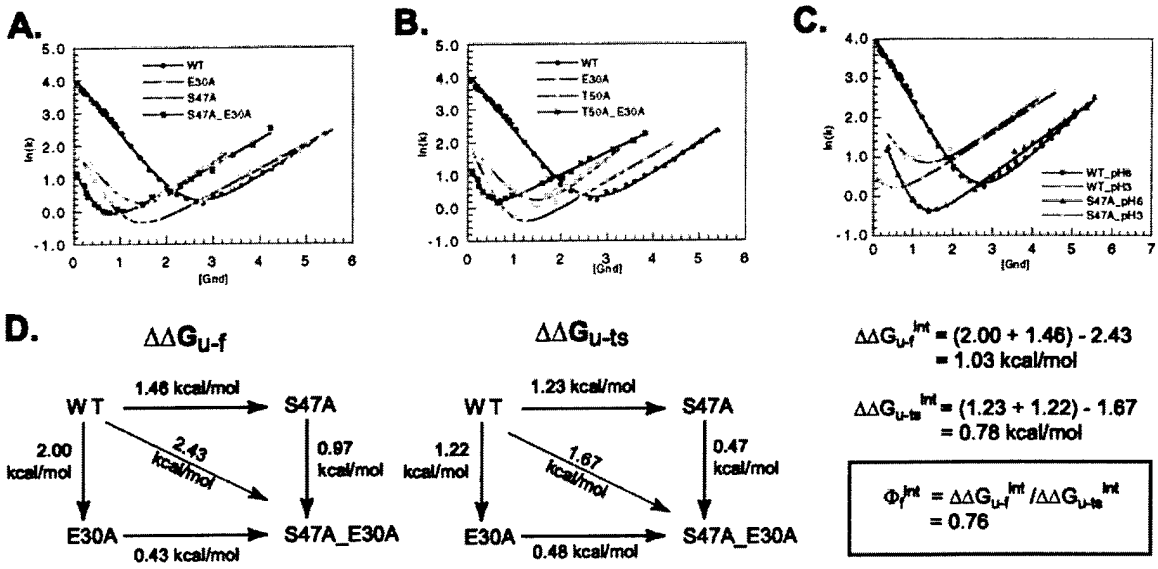


Fig. 2. Kinetic analysis of E30A_S47A (*A*) and E30A_T50A (*B*) in 50 mM NaPi, pH 6, at 295 K. Gnd, guanidine. (*C*) Kinetic analysis of WT and S47A mutant in 50 mM NaPi, pH 3, at 295 K. Solid lines indicate the best fit to the data with KALEIDAGRAPH. (*D*) Double-mutant cycle analysis to determine the interaction energy between E30A and S47A in the native ($\Delta\Delta G_{U\text{-}F}$) and transition states ($\Delta\Delta G_{U\text{-}\ddagger}$). Data for the calculations are presented in Table 1. $\Delta\Delta G_{U\text{-}\ddagger}$ is calculated as $RT\ln(k_f^{wt}/k_f^{mut})$. Energies for the E30A_T50A double mutant are not shown explicitly but can be calculated similarly by using data from Table 1.

Backbone Restriction in the Folding Transition State. In addition to manipulating interaction energies of side chains by point mutagenesis, we can alter the entropy of the protein backbone by engineering loop insertions and covalent crosslinks and then use the effect of such changes on the folding kinetics as a reporter on the degree of association of the structural elements in the transition state. As in Φ-value analysis, the premise of these experiments is that changes in the rates of folding and unfolding will reveal the extent to which two elements are brought together in the transition state compared with the denatured and native states. To facilitate interpretation of the kinetic results, it is reasonable to assume that the primary effect of these modifications is on chain conformational entropy. The denatured state, which is most disordered, is expected to be affected significantly by the modifications: glycine insertion in loops increases the entropy of the denatured state and thus lowers its free energy, whereas disulfide crosslinks decrease the entropy of the denatured state and destabilize it in proportion to the length of the crosslinked fragment (Fig. 3). The entropy of the native state, on the other hand, should not change greatly in the case of the disulfide crosslinks, because the elements we are probing already interact fully in the native state. In the case of the 10-glycine mutant, the entropy of the native state will increase but to a lesser extent than the entropy of the denatured state. There might be some destabilization of the native state resulting from disruption of local interactions at the site of glycine insertion or strain from suboptimal disulfide geometry; however, we have tried to minimize these effects by choosing the modification sites carefully. The entropic effect on the transition state then depends on the proximity of the structural elements being probed and can be deduced from the changes in k_f and k_u. If only k_f is affected, it can be concluded that the crosslinked regions are as ordered in the transition state as in the native state (Fig. 3 *A* and *B*), whereas, if only k_u changes, the region is likely to be as disordered in the transition state as in the denatured state.

Table 1. Kinetic parameters for the hydrogen bond network mutants

Mutant	$\ln(k_f)^{0M}$	$\ln(k_u)^{3M}$	m_f	m_u	$\Delta G_{U\text{-}F}$	$\Delta\Delta G_{U\text{-}F}$
WT*	4.10	0.139	1.02	0.54	3.95	—
E30A*	2.03	1.48	1.09	0.65	2.28	−2.00
S47A*	2.01	0.537	1.50	0.44	2.18	−1.46
T50A*	1.99	0.750	1.84	0.47	2.14	−1.60
E30A_S47A	1.21	1.39	†	0.46	1.28	−2.43
E30A_T50A	1.32	1.66	†	0.39	0.977	−2.52
WT_pH3	2.08	1.69	1.08	0.41	1.46	−2.09
S47A_pH3	0.052	1.59	1.45	0.38	0.249	−3.23

Kinetics of folding and unfolding were followed by changes in tryptophan fluorescence on a stopped flow instrument at 295 K; k_f is reported in the absence of denaturant, and k_u is in 3 M Gnd to avoid extrapolation; m_f and m_u are the dependences of the folding and the unfolding rates, respectively, on Gnd. $\Delta G_{U\text{-}F}$ (free energy of unfolding) and $\Delta\Delta G_{U\text{-}F}$ (the difference in $\Delta G_{U\text{-}F}$ between WT and the mutant proteins) were calculated from the kinetic parameters as described in the *Methods* section. Typical errors for the kinetic measurements are 1–10% as reported in ref. 13.

*Kinetic data for these mutants were published previously in ref. 12.

†These values could not be estimated reliably because of the small region over which k_f can be measured.

Glycine Insertion in the n-src Loop. The involvement of the n-src loop in the transition state was difficult to evaluate from the point mutagenesis analysis because of the very small effect on stability of all of the mutations in this region (13). An insertion of 10 glycine residues in the n-src loop between residues 40 and 41 (Fig. 3*A*) was designed to test the importance of this loop in bringing the distal β-hairpin and the diverging turn together at the transition state. The addition of 10 glycine residues considerably increases the entropic cost of bringing these two elements together and is expected to decrease the rate of folding if association of the two elements is required in the transition state. A correlation between the rate of folding and the length of this loop was observed in a comparison of homologous SH3 domains (17): the phosphatidylinositol 3-kinase SH3 domain has the longest n-src loop, and its folding rate is the slowest of the SH3 domains that have been characterized (18). Furthermore, in a

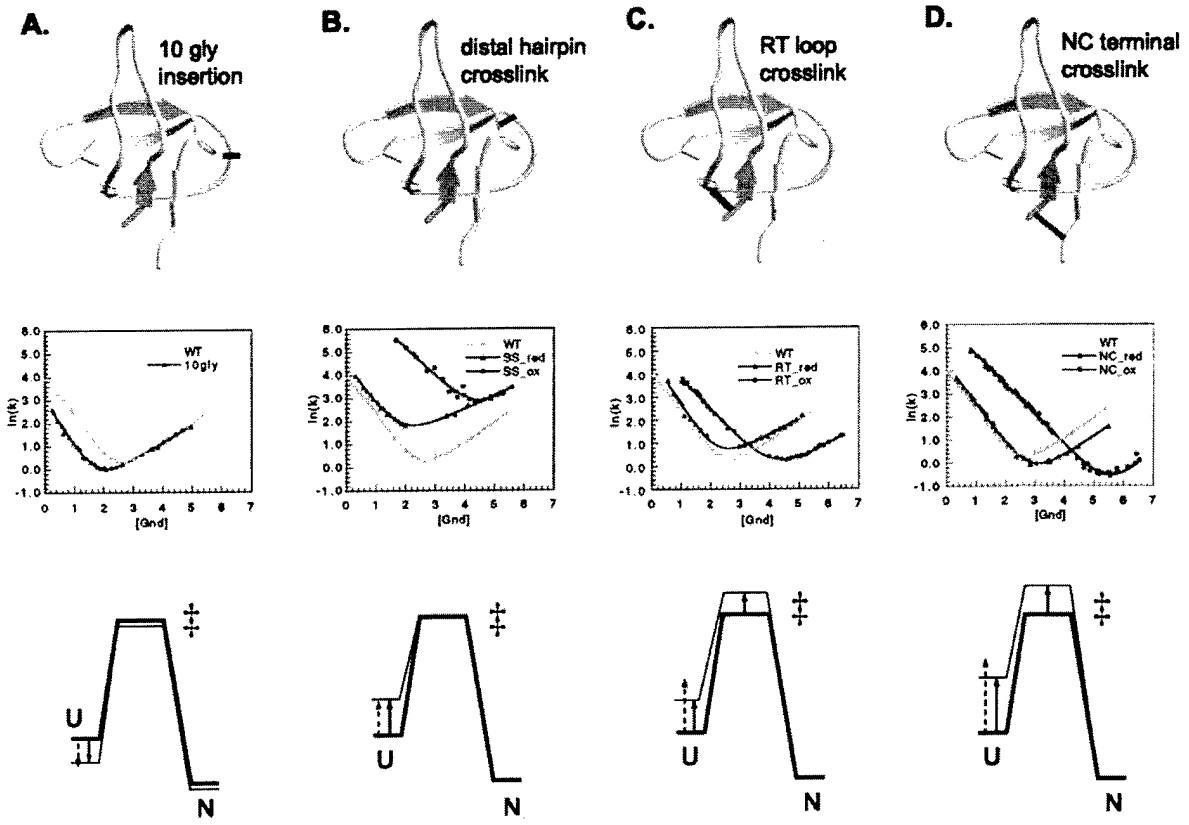


Fig. 3. Characterization of the 10-glycine insertion (*A*), the distal hairpin crosslink (*B*), the RT loop crosslink (*C*), and the N- and C-terminal crosslink mutants (*D*). For each modification, panels from *Top* to *Bottom* show site of the modification, kinetic analysis, and an energy diagram explaining the effects on kinetics. The structure diagrams of the SH3 domain were generated with MOLSCRIPT (33) and RASTER3D (34, 35). Energy diagrams show the relative free energies of the unfolded (U), the transition (‡), and the native state (N) before (thick lines) and after (thin lines) each modification. In the case of the 10-glycine insertion, the reference protein is the WT, and, for the three disulfide crosslinks, the reference is the reduced version of the particular double cysteine mutant. Solid arrows indicate the experimentally determined changes in the free energy of denatured state, and dashed arrows indicate the changes predicted from loop entropy estimates (16). 10gly, 10-glycine; NC, N- and C-terminal crosslink; SS, distal hairpin crosslink; red, reduced; ox, oxidized.

combinatorial mutagenesis experiment on the SH3 domain, phage-display selection of correctly folded proteins yielded an enrichment for shortened n-src loops (19). The 10-glycine insertion leads to an overall decrease in stability of 0.72 kcal/mol (Table 2), consistent with loop entropy estimates ($-1.5RT\ln(L/L_o) = 0.78$ kcal/mol, where L_o and L are the original and the new loop lengths, respectively; ref. 16). Remarkably, the kinetic effect of the insertion (Fig. 3*A* and Table 2) is exclusively on the rate of folding, k_f. The structural elements on either end of the n-src loop (the distal β-hairpin and the diverging turn) seem to be fully associated at the folding transition state, consistent with the double-mutant results (Fig. 2). The lack of effect of such a large insertion on the unfolding rate, k_u, is striking and suggests that unfolding is initiated by the dissociation of the two terminal strands and the RT loop but does not involve disruption of the n-src loop.

Crosslinking the Distal β-Hairpin. Φ-Value analysis of both the src and the α-spectrin SH3 domains highlighted the importance of the distal β-hairpin in the folding transition state (12, 20). For both proteins, the tip of the distal β-hairpin is the only region that contains residues with Φ_F-values equal to 1 (in the SH3 domain, S47, T50, and G51 have all of their local interactions formed at the transition state). The middle of the hairpin, however, appeared somewhat flexible (L44A and Y55A on the solvent-exposed side of the hairpin have intermediate Φ-values), and there were no suitable mutations to probe strand association at the base of the hairpin. To examine specifically the backbone

Table 2. Kinetic parameters for 10-glycine insertion and disulfide crosslink mutants

Mutant	$\ln(k_f)^{1M}$	$\ln(k_u)^{5.5M}$	m_f	m_u	ΔG_{U-F}	$\Delta\Delta G_{U-F}$
WT	2.36	2.44	1.02	0.54	3.95	—
10gly	0.997	2.30	1.20	0.44	2.86	−0.72
SS_red	2.75	3.38	1.05	0.33	2.50	−0.32
SS_ox	6.50	3.37	0.750	0.59	5.83	1.88
RT_red	2.65	2.37	1.18	0.39	3.47	0.21
RT_ox	3.81	0.62	0.78	0.42	4.98	1.92
NC_red	2.58	1.59	1.03	0.50	4.36	0.63
NC_ox	4.73	−1.30	0.766	0.80	8.70	3.58

Kinetics of folding and unfolding were followed by changes in tryptophan fluorescence on a stopped flow instrument at 295 K; k_f is reported in 1 M Gnd, and k_u is in 5.5 M Gnd to avoid extrapolation; m_f and m_u are the dependences of the folding and the unfolding rates, respectively, on Gnd. ΔG_{U-F} (free energy of unfolding) and $\Delta\Delta G_{U-F}$ (the difference in ΔG_{U-F} between the particular mutant protein and WT) were calculated from the kinetic parameters as described in the *Methods* section. Typical errors for the kinetic measurements are 1–10% as reported in ref. 13. Abbreviations are defined in the legend to Fig. 3.

conformation of the β-hairpin in the transition state, we tested the kinetic consequences of covalently crosslinking the hairpin. For this purpose, W43 and S58 at the base of the distal β-hairpin (Fig. 3*B*) were mutated to cysteines and then crosslinked by using an oxidizing agent (see *Methods* section). Under reducing conditions, the double-cysteine mutant (SS mutant) is destabilized compared with the WT SH3 domain; however, oxidation significantly stabilizes the mutant protein: $\Delta\Delta G_{U\text{-}F}$ between the reduced and oxidized forms of the mutant is 2.55 kcal/mol. This value matches closely the theoretical loop entropy estimate from polymer theory (2.38 kcal/mol; see *Methods* section; ref. 16), suggesting that stabilization results largely from the decrease in entropy of the denatured state. Kinetic analysis reveals that the oxidized protein folds 30 times faster than the reduced form, whereas the unfolding rate is virtually unchanged (Fig. 3*B* and Table 2). The resulting Φ_F-value of 1 for the disulfide crosslink unambiguously confirms that the distal β-hairpin is conformationally restricted at the transition state (Fig. 3*B*). As in the case of the glycine insertion, the lack of effect of the crosslink on k_u is remarkable: the unfolding event must not involve even partial unraveling of the distal β-hairpin. Flexibility in the middle of the hairpin therefore does not prevent the conformational locking of the hairpin's base.

Crosslinking the RT Loop. To probe the extent of formation of the RT loop in the folding transition state, we have introduced a disulfide crosslink at its base. T9 at the N terminus of the protein and Q33 after the diverging turn were mutated to cysteine and oxidized to close off a loop of 23 residues covalently (Fig. 3*C*). Oxidation stabilizes the protein by 1.71 kcal/mol, suggesting that the disulfide bond geometry is favorable and does not introduce strain in the native state of the protein. This stabilization, however, is less than the loop entropy reduction estimate (2.75 kcal/mol; see *Methods* section; ref. 16), suggesting that the RT loop might be partially structured in the denatured state. Kinetic analysis (Fig. 3*C* and Table 2) shows that, in contrast to the distal hairpin crosslink, formation of the RT loop crosslink dramatically decreases the unfolding rate, suggesting that the rate-limiting step in unfolding involves the opening of the RT loop. Crosslinking also increases the folding rate, indicating that parts of the RT loop might be structured in the transition state. Considering the point mutagenesis results that revealed Φ-values uniformly close to 0 throughout the N-terminal strand and the tip of the RT loop, it is most likely that the crosslink itself has caused an expansion of the structured region of the transition state to include the RT loop. Another possibility is that the RT loop is stabilized primarily by backbone hydrogen bonds and not by side chain–side chain interactions in the transition state.

Crosslinking the N and C Termini. The point mutagenesis analysis suggested that the N and C termini do not associate at the folding transition state, because none of the mutations in this region affect the rate of folding. We probed the association between the N- and C-terminal strands in the transition state by engineering a disulfide crosslink between them. Two cysteine mutations were introduced in the SH3 domain (T9C at the N terminus and S64C at the C terminus) and then crosslinked as described for the distal hairpin. Comparison of the T9C_S64C mutant (NC mutant) under reducing and oxidizing conditions (Fig. 3*D* and Table 2) reveals that the oxidized protein is stabilized significantly ($\Delta\Delta G_{U\text{-}F}$ = 2.8 kcal/mol) but less than expected from the effect of crosslinking on the entropy of the denatured state ($\Delta\Delta G$ = 3.52 kcal/mol; see *Methods* section; ref. 16; the denatured state may be more ordered than the random coil model assumed in the loop entropy estimate). Kinetic measurements show that both the folding and the unfolding rates of the SH3 domain are affected roughly equally by the disulfide crosslink. In general, circularization is always expected to increase k_f because of the greater decrease in the entropy of the denatured state compared with that of the transition state, but k_u will decrease only if the termini are apart at the transition state. (In proteins, like acyl-CoA-binding protein, whose termini interact at the transition state, crosslinking would be expected to affect primarily the rate of folding; ref. 7.) The decrease in the unfolding rate brought about by the NC crosslink suggests that the termini are not as ordered in the SH3 transition state as they are in the native state.

Discussion

SH3 Folding Transition State. Φ-Value analysis has become the method of choice for studying folding transition states (4, 5). In this study, we have extended the repertoire of probes of the transition state ensemble to include glycine insertion and disulfide crosslinking as direct experimental measures of the conformational constraints on the peptide backbone and the association of structural elements. Engineering of disulfide crosslinks has been used to assess the effect of reducing conformational entropy in different parts of the molecule (21–25) or to explore transition-state heterogeneity (26, 27). Loop insertions have been used to explore the role of loops in determining protein stability and the folding mechanism (28–30).

Our current findings from the double-mutant analysis, 10-glycine insertion, and covalent crosslinking combined with the point mutagenesis results provide a comprehensive picture of the folding transition state for the SH3 domain. The distal β-hairpin still stands out as the structural element best formed in the transition state; however, now we have information about strand pairing along the entire length of the hairpin. Previous experiments had indicated that the tip of the hairpin is well ordered as judged by the clustering of high Φ-value residues; however, the middle of the hairpin is likely not as rigid in the transition state as in the native state, because solvent-exposed residues are paired only partially with each other. The finding that the distal β-hairpin crosslink increases the rate of folding and has no effect on the unfolding rate suggests that the two strands come in close proximity at the base of the hairpin before the folding transition state. Interestingly, similar results were found for one of the hairpins in protein L (38), suggesting that "looping" in the middle of the hairpin might be a common theme during protein folding. Constraining the tip and the base of the hairpin may be important for specifying the topology of the folding protein, whereas keeping the middle flexible would allow hydrophobic core rearrangements after the transition state.

The current experiments establish firmly the interaction of the distal β-hairpin and the diverging turn in the transition state. The double-mutant results provide concrete evidence that the hydrogen bond network between these two elements is mostly formed at the transition state. The high Φ_F-values for these nonlocal hydrogen bond interactions (0.78 and 0.83) indicate that most of the interaction energy is present at the rate-limiting step; further alignment of the hydrogen bond geometries and immobilization of the participating residues after the transition state must contribute the remaining ≈20%. It is generally assumed that transition states for folding are stabilized primarily by hydrophobic interactions, with hydrogen bonds contributing only later to the stability of the native state because of their stricter geometric requirements. The SH3 domain is the first case in which nonlocal side chain hydrogen bonds have been found to stabilize the transition state (31). The hydrogen bond network does not, however, seem to be required for folding (e.g., to confer specificity or determine the alignment of structural elements). In the α-spectrin SH3 domain, for example, this interaction is replaced by a hydrophobic cluster, and in the fyn SH3 domain (78% homologous to the SH3 domain), the same hydrogen bond network is in place in the native state but does not contribute to stabilization of the transition state (A. Davidson, personal communication). The difference between the src and fyn SH3

domains may be a result of the slightly different structures of their diverging turns: perhaps the two large phenylalanines in fyn (instead of Phe and Leu in src) cannot pack closely next to each other and interact with the distal β-hairpin until after the transition state. The local variation in the types of interactions stabilizing the transition state is an illustration of the lack of selective pressure on the details of the protein-folding mechanism (32).

The association of the distal β-hairpin and the diverging turn at the transition state is confirmed further by the 10-glycine insertion into the n-src loop connecting the two elements. The exclusive effect of the insertion on the folding rate strongly suggests that the peptide backbone of the regions flanking the loop is constricted at the rate-limiting step. In marked contrast, all previous loop-insertion experiments revealed that both folding rates decrease and unfolding rates increase as loops lengthen. The context dependence of loop lengthening has been noted before; however, it has been explained mostly in terms of the flexibility of the region in the native state (28–30). Our findings indicate that the kinetic effect of loop elongation is related directly to the extent to which the elements connected by the loop are topologically constrained at the transition state.

The effects of the disulfide crosslinks and the loop insertion on the unfolding rate highlight the pronounced hierarchy of events. The distal β-hairpin crosslink and the glycine loop insertion have no effect on k_u, clearly indicating that the three-stranded sheet formed by the interaction of the distal β-hairpin and the diverging turn remains intact at the unfolding transition state. In contrast, the RT loop crosslink and the NC crosslink dramatically slow down the unfolding rate, suggesting that the rate-limiting step in unfolding involves the dissociation of the N and C termini and the opening of the RT loop. These results indicate an even greater structural polarization of the folding transition state than suggested by our previous studies.

The conformational restriction of structural elements in the src SH3 domain transition state has implications for the mechanism of folding. The transition state ensemble consists of a relatively small number of conformers, all of which have the distal β-hairpin and the diverging turn ordered and interacting with each other, and with heterogeneities limited to the RT loop and the N and C termini. The energy landscape of this protein, therefore, significantly deviates from a symmetrical funnel in which the transition state includes all conformations with a particular degree of freedom. The folding of the src SH3 domain is surprisingly consistent with the more traditional single pathway-based picture of protein folding.

We thank Andreas Matouschek for providing us with his disulfide-crosslinking protocol. We are grateful to members of the Baker group for their useful comments on the manuscript. This work was supported by grants from the National Institutes of Health and the Office of Naval Research and a Young Investigator award to D.B. from the Packard Foundation.

1. Pande, V. S., Grosberg, A., Tanaka, T. & Rokhsar, D. S. (1998) *Curr. Opin. Struct. Biol.* **8,** 68–79.
2. Shakhnovich, E. I. (1998) *Folding Des.* **3,** R108–R111.
3. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994) *Biochemistry* **33,** 10026–10036.
4. Fersht, A. R., Matouschek, A. & Serrano, L. (1992) *J. Mol. Biol.* **224,** 771–782.
5. Fersht, A. R. (1995) *Curr. Opin. Struct. Biol.* **5,** 79–84.
6. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999) *Nat. Struct. Biol.* **6,** 1005–1009.
7. Kragelund, B. B., Osmark, P., Neergaard, T. B., Schiodt, J., Kristiansen, K., Knudsen, J. & Poulsen, F. M. (1999) *Nat. Struct. Biol.* **6,** 594–601.
8. López-Hernández, E. & Serrano, L. (1995) *Folding Des.* **1,** 43–55.
9. Main, E. R., Fulton, K. F. & Jackson, S. E. (1999) *J. Mol. Biol.* **291,** 429–444.
10. Martinez, J. C. & Serrano, L. (1999) *Nat. Struct. Biol.* **6,** 1010–1016.
11. Milla, M. E., Brown, B. M., Waldburger, C. D. & Sauer, R. T. (1995) *Biochemistry* **34,** 13914–13919.
12. Grantcharova, V. P., Riddle, D. S., Santiago, J. V. & Baker, D. (1998) *Nat. Struct. Biol.* **5,** 714–720.
13. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. & Baker, D. (1999) *Nat. Struct. Biol.* **6,** 1016–1024.
14. Thornton, J. M. (1981) *J. Mol. Biol.* **151,** 261–287.
15. Grantcharova, V. P. & Baker, D. (1997) *Biochemistry* **36,** 15685–15692.
16. Jacobsen, H. & Stockmayer, W. H. (1950) *J. Chem. Phys.* **18,** 1600–1606.
17. Plaxco, K. W., Guijarro, J. I., Morton, C. J., Pitkeathly, M., Campbell, I. D. & Dobson, C. M. (1998) *Biochemistry* **37,** 2529–2537.
18. Guijarro, J. I., Morton, C. J., Plaxco, K. W., Campbell, I. D. & Dobson, C. M. (1998) *J. Mol. Biol.* **276,** 657–667.
19. Riddle, D. S., Santiago, J. V., Bray-Hall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q. & Baker, D. (1997) *Nat. Struct. Biol.* **4,** 805–809.
20. Martinez, J. C., Pisabarro, M. T. & Serrano, L. (1998) *Nat. Struct. Biol.* **5,** 721–729.
21. Creighton, T. E. (1992) in *Protein Folding*, ed. Creighton, T. E. (Freeman, New York), pp. 301–354.
22. Strausberg, S., Alexander, P., Wang, L., Gallagher, T., Gilliland, G. & Bryan, P. (1993) *Biochemistry* **32,** 10371–10377.
23. Ikeguchi, M., Fujino, M., Kato, M., Kuwajima, K. & Sugai, S. (1998) *Protein Sci.* **7,** 1564–1574.
24. Kobayashi, N., Honda, S. & Munekata, E. (1999) *Biochemistry* **38,** 3228–3234.
25. Zhang, T., Bertelsen, E. & Alber, T. (1994) *Nat. Struct. Biol.* **1,** 434–438.
26. Moran, L. B., Schneider, J. P., Kentsis, A., Reddy, G. A. & Sosnick, T. R. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 10699–10704.
27. Otzen, D. E. & Fersht, A. R. (1998) *Biochemistry* **37,** 8139–8146.
28. Viguera, A. R. & Serrano, L. (1997) *Nat. Struct. Biol.* **4,** 939–946.
29. Ladurner, A. G. & Fersht, A. R. (1997) *J. Mol. Biol.* **273,** 330–337.
30. Nagi, A. D., Anderson, K. S. & Regan, L. (1999) *J. Mol. Biol.* **286,** 257–265.
31. Myers, J. K. & Oas, T. G. (1999) *Biochemistry* **38,** 6761–6768.
32. Plaxco, K. W., Simons, K. T. & Baker, D. (1998) *J. Mol. Biol.* **277,** 985–994.
33. Kraulis, P. J. (1991) *J. Appl. Crystallogr.* **24,** 946–950.
34. Bacon, D. J. & Anderson, W. F. (1988) *J. Mol. Graphics* **6,** 219–220.
35. Merritt, E. A. & Bacon, D. J. (1997) *Methods Enzymol.* **277,** 505–524.
36. Ferrin, T. E., Huang, C. C., Jarvis, L. E. & Langridge, R. (1988) *J. Mol. Graphics* **6,** 13–27.
37. Huang, C. C., Pettersen, E. F., Klein, T. E., Ferrin, T. E. & Langridge, R. (1991) *J. Mol. Graphics* **9,** 230–236.
38. Kim, D. E., Fisher, C. & Baker, D. (2000) *J. Mol. Biol.* **298,** 971–984.

doi:10.1006/jmbi.2000.3816 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 299, 1341–1351

JMB

NMR Characterization of Residual Structure in the Denatured State of Protein L

Qian Yi, Michelle L. Scalley-Kim, Eric J. Alm and David Baker*

Department of Biochemistry
University of Washington
Seattle, WA 98195, USA

Triple-resonance NMR experiments were used to assign the $^{13}C^{\alpha}$, $^{13}C^{\beta}$, ^{15}N and NH resonances for all the residues in the denatured state of a destabilized protein L variant in 2 M guanidine. The chemical shifts of most resonances were very close to their random coil values. Significant deviations were observed for G22, L38 and K39; increasing the denaturant concentration shifted the chemical shifts of these residues towards theory random coil values. Medium-range nuclear Overhauser enhancements were detected in segments corresponding to the turn between the first two strands, the end of the second strand through the turn between the second strand and the helix, and the turn between the helix and the third strand in 3D H^1, N^{15}-HSQC-NOESY-HSQC experiments on perdeuterated samples. Longer-range interactions were probed by measuring the paramagnetic relaxation enhancement produced by nitroxide spin labels introduced *via* cysteine residues at five sites around the molecule. Damped oscillations in the magnitude of the paramagnetic relaxation enhancement as a function of distance along the sequence suggested native-like chain reversals in the same three turn regions. The more extensive interactions within the region corresponding to the first β-turn than in the region corresponding to the second β-turn suggests that the asymmetry in the folding reaction evident in previous studies of the protein L folding transition state is already established in the denatured state.

Keywords: protein L; denatured state; residual structure; paramagnetic relaxation enhancement; NMR

*Corresponding author

Introduction

How a protein folds into a unique three-dimensional structure is one of the greatest questions of modern structural biology. An exciting recent development is the direct structural characterization of the starting point of the folding reaction, the denatured state, using multi-dimensional NMR methods. It has been found (Neri *et al.*, 1992; Logan *et al.*, 1994; Shortle 1996a,b; Eliezer *et al.*, 1998; Fong *et al.*, 1998; Mok *et al.*, 1998) that the unfolded states of several proteins under both denaturing and native solution conditions can contain a significant amount of residual structure. Recent studies under non-denaturing conditions have demonstrated that the denatured state ensemble of staphylococcal nuclease has a native-like overall topology (Gillespie & Shortle, 1997), while the overall topology of the drk SH3 unfolded states is not native-like, although it has some native-like features (Mok *et al.*, 1998). The detailed structural characterization of the denatured states of proteins whose folding transition has been extensively studied should increase understanding of the early stages of the folding process.

We have chosen the IgG binding domain of protein L as a model system for understanding folding in detail. The folding of protein L has been characterized using a wide range of methods (Yi & Baker, 1996; Scalley *et al.*, 1997; Yi *et al.*, 1997; Gu *et al.*, 1997; Plaxco *et al.*, 1999; Kim *et al.*, 1999) and the folding transition state has been extensively mapped through the analysis of the effect of 70 point mutants distributed around the protein. A destabilized mutant (F20W/Y32A) of protein L has recently been characterized by circular dichroism and stopped-flow kinetics (Scalley *et al.*, 1999).

Abbreviations used: PRE, paramagnetic relaxation enhancement; HSQC, heteronuclear single quantum coherence; NOESY, NOE spectroscopy; CD, circular dichroism.

E-mail address of the corresponding author: dabaker@u.washington.edu

0022-2836/00/051341–11 $35.00/0

These data suggested the presence of residual structure in 2 M to 3 M guanidinium chloride. In this study we use NMR methods to gain more specific structural information on this denatured state of protein L.

Results

Assignment of the backbone $^{13}C^{\alpha}$, $^{13}C^{\beta}$, ^{15}N and NH resonances of F20W/Y32A

Characterization of denatured states of proteins using NMR techniques is often challenging, since the chemical shift dispersion of most resonances is poor because of conformational averaging. However, the backbone ^{15}N and ^{13}CO chemical shifts, which are mainly influenced by residue type and the local amino acid sequence (Braun *et al.*, 1994; Yao *et al.*, 1997), remain well dispersed in the denatured state. Using ^{13}C, ^{15}N-double labeled protein sample and triple resonance NMR experiments (see Materials and Methods), we have assigned the $^{13}C^{\alpha}$, $^{13}C^{\beta}$, ^{15}N and NH resonances for all the residues of the F20W/Y32A mutant in 2 M guanidine at pH 5.0.

$^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts are predominantly determined by backbone conformation, and the perturbations of the $^{13}C^{\alpha}$ and $^{13}C_{\beta}$ chemical shifts from their random coil values (Spera & Bax, 1991; Wishart & Skyes, 1994) are reliable indicators of secondary structure in folded proteins. In general, $^{13}C^{\alpha}$ resonances are shifted downfield by an average of 2.6 ppm for α-helices, and shifted upfield by 1.7 ppm for β-sheets. Figure 1 shows that the deviations of the $^{13}C^{\alpha}$ chemical shifts of F20W/Y32A in 2 M guanidine at pH 5.0 from the random coil values are very small, indicating the population of regular secondary structure is very low under these conditions. However, small but consistent upfield chemical shift perturbations were observed for every residue of the segment from A31 to D41 (helical in native protein L structure), suggesting that there is some residual helical content in F20W/Y32A in 2 M guanidine. Also, small, but consistent downfield chemical shift perturbations were observed for every residue from Y45 to A50 (the third strand in native protein L structure), suggesting that this segment has some preference for extended conformations under these conditions.

Previous chemical denaturation studies using fluorescence techniques suggested that there may be some residual structure around W20 in 2-3 M guanidine that is lost at higher (>3 M) guanidine concentrations (Scalley *et al.*, 1999). A guanidine titration ranging from 1.5 M to 5 M guanidine was carried out to investigate possible conformational changes in the denatured state ensemble. There is no dramatic change in the ^{1}H-^{15}N HSQC spectra from 1.5 M to 5 M guanidine except that small, but significant chemical shift perturbations (~0.10 ppm shifting toward the random coil values) were observed for the amide group protons of G22, L38 and K39 (Figure 2(a)). These shifts are roughly linear functions of the denaturant concentration; there is little indication of a cooperative transition between different populations (Figure 2(b)). These results suggest that there is some residual structure around G22, L38 and K39 in denatured F20W/Y32A in 2-3 M guanidine.

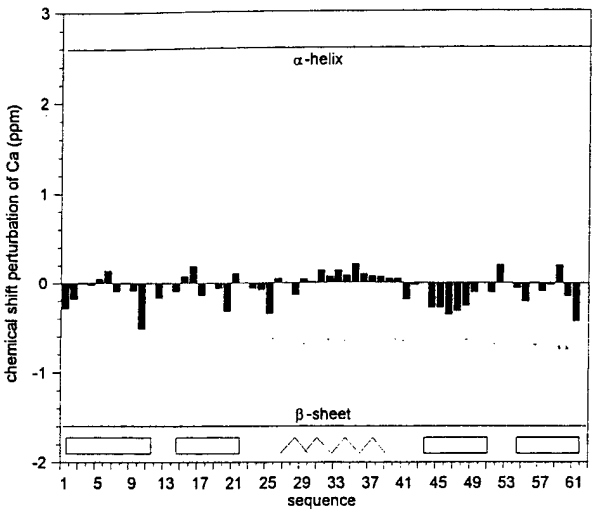


Figure 1. Chemical shift perturbations of the $^{13}C^{\alpha}$ resonances of F20W/Y32A in 2 M guanidine from the random coil values. The two straight horizontal lines at 2.6 ppm and −1.7 ppm represent the perturbations expected for regular α-helix and β-sheet conformations respectively. The random coil values were taken from Wishart & Sykes, 1994.

Nuclear Overhauser enhancements (NOEs) between amide protons

NOEs between amide group protons were obtained using 3D H^1, N^{15}-HSQC-NOESY-HSQC experiments on perdeuterated F20W/Y32A in 2-3 M guanidine. Recent work on the drk SH3 domain demonstrated that deuteration greatly enhances NOESY-based studies of denatured proteins because of longer relaxation time due to the reduced spin diffusion (Sattler & Fesik 1996). Sequential $(i, i+1)$ and $(i, i+2)$ HN-HN NOEs were observed for most residues of F20W/Y32A in 2 M guanidine. Only a few medium-range $(i, i+3)$ and $(i, i+4)$ HN-HN NOEs were detected (Figure 3). All observed medium-range NOEs are located in segments close to turn regions in native protein L. These segments correspond to the first hairpin turn, the end of the second β-strand before the helix and the turn following the helix. Thus, there are some native-like turn structures populated in the denatured states of F20W/Y32A in 2-3 M guanidine. This is consistent with the result from guanidine titration experiments described above. NOEs in the region corresponding to the second β-turn in native protein L were conspicuously absent.

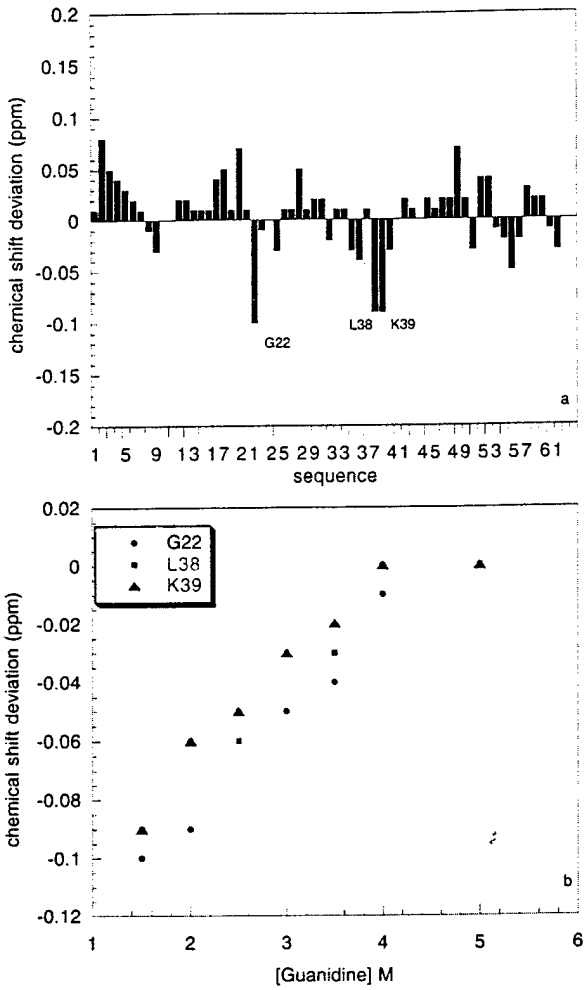


Figure 2. Guanidine-dependence of amide proton chemical shifts. (a) Deviation of chemical shifts in 1.5 M from values in 5.0 M guanidine for all the amide group protons of F20W/Y32A/C63. (b) Guanidine dependence of the amide group proton chemical shifts of G22, L38 and K39.

Measurement of paramagnetic relaxation enhancement (PRE) by nitroxide spin label

To probe longer-range interactions, we examined the paramagnetic relaxation enhancement of the amide group protons due to introduced nitroxide spin labels. This technique has been used successfully to characterize the denatured state of staphylococcal nuclease under native conditions (Gillespie & Shortle, 1997). The advantage of PRE is that the free electron from the nitroxide label increases the relaxation rate of protons over a distance of up to 20 to 25 Å. In contrast, the NOE between two protons only extends up to 10 Å even at very long mixing times (Mok *et al.*, 1998). Thus, PRE can be useful for studying weak long-range molecular interactions in relatively disordered denatured states.

Paramagnetic relaxation enhancement increases relaxation rates in a distance-dependent manner. The enhancement effect is described by the Solomon-Bloembergen equations (Solomon & Bloembergen, 1956; Kosen 1989):

$$\Delta(1/T_1) = \Delta R_1 = 2K(3\tau_c/(1+\omega_H^2\tau_c^2))/r^6 \quad (1)$$

$$\Delta(1/T_2) = \Delta R_2 = K(4\tau_c + 3\tau_c/(1+\omega_H^2\tau_c^2))/r^6 \quad (2)$$

where K is 1.23×10^{-32} cm^6 s^{-2} for a nitroxide radical, r is the distance between the electron and the proton, τ_c is the correlation time for the electron-proton vector, and ω_H is the Larmour frequency of the proton. Equations (1) and (2) are based on the assumptions that the vector between the electron and the proton is free to undergo isotropic rotational diffusion, and that its length is fixed. Both equations are valid only for relaxation due to the magnetic interaction between a single unpaired electron and a proton of a macromolecule when τ_c is greater than 10^{-9} second and ω_H is between 400 and 600 MHz.

To provide attachment sites for the spin labels, cysteine residues were introduced one at a time at strategic locations on the surface of protein L. To minimize perturbations to the structure accompanying the cysteine substitutions, the residues at the chosen positions were highly solvent-accessible and make few interactions with other residues. To obtain information on all parts of the protein, one probe was introduced into each of the five secondary structural elements in the protein. The sites at which cysteine residues were introduced are shown in Figure 4; E1C is at the N terminus, T17C in the middle of the second strand, S29C in the middle of the helix, T46C in the middle of the third strand and C63 was added at the C terminus. Nitroxide spin labels were attached to the introduced cysteine residues as described in Materials and Methods.

Direct measurement of the effect of the paramagnetic relaxation enhancement on T_1 and T_2 can be carried out using standard NMR techniques, such as the inversion-recovery sequence and CPMG spin echo sequence, but these techniques can be quite time-consuming. Instead, it is more efficient to measure the decrease of ^{1}H-^{15}N peak intensity in HSQC spectra. The peak linewidth in the HSQC spectrum is increased due to faster transverse relaxation (R_2) of ^{1}H, and ultimately the peak intensity is decreased. Figure 5 shows the PRE effect on peak intensity by comparing ^{1}H-^{15}N HSQC spectra of T17C* with and without the unpaired free electron present (oxidized and reduced forms respectively). The magnitude of the PRE for each residue in each protein was determined using the simulation method described by Gillespie and co-workers (Gillespie & Shortle, 1997) (see Materials and Methods) and is shown in Figure 6. The bars indicate the magnitude of the PRE, the arrows indicate the position of the label,

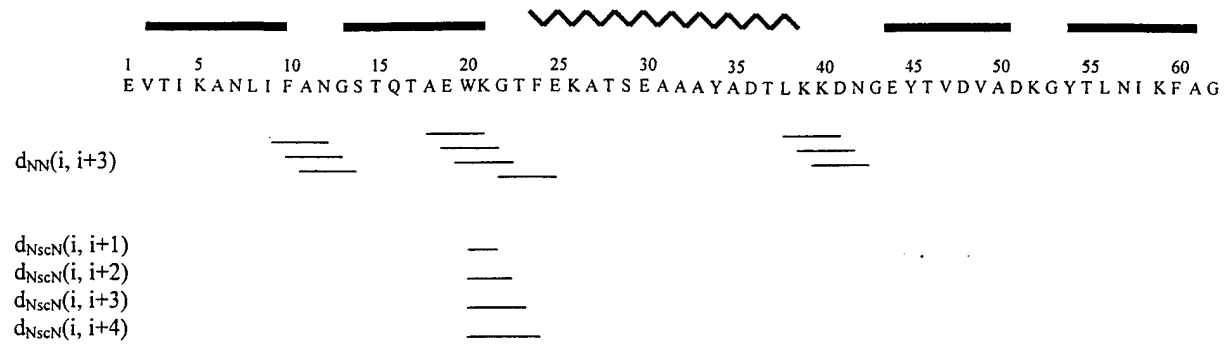


Figure 3. Schematic of the medium-range HN-HN NOEs observed in HSQC-NOESY-HSQC experiments with a 600 ms mixing time along with the protein sequence and the secondary structural elements of the native protein L. The filled bars represent β-strands and the zigzag line represents α-helix. The d_{NscN} NOEs refer to the NOEs observed between the side-chain NεH of W20 and the backbone amide group protons.

and the boxes at the top indicate the positions of the secondary structure elements. For a number of residues (indicated by the hatched bars) close to the introduced labels, the relaxation enhancement was too great to be measured.

To facilitate interpretation of the PRE results, we have compared the experimental PRE profiles with profiles generated by computer simulation for: (i) the native state; (ii) a model of the denatured state of protein L in which local sequence-structure relationships are preserved; and (iii) a random chain model of the protein L denatured state (see Materials and Methods). The PRE profile expected for an ensemble of configurations without any persistent structure should smoothly decrease with increasing sequence distance from the introduced probe. Such a smooth decrease is clearly evident in the random chain simulations (Figure 7, row D). The experimental PRE profiles in all five cysteine mutants (Figure 7, row B) show the expected decrease in relaxation enhancement with distance from the spin label (indicated by the arrows in Figure 6). Superimposed on the gradual decay of relaxation enhancement effect are oscillations not seen in the random chain model that suggest the presence of chain reversals. The differences between the profiles in rows B and D in Figure 7 are likely to be due to residual structure in the protein L denatured state. For example, for the E1C* mutant, residues from N12 to T15, which are closer to the nitroxide probe in the sequence, experienced less broadening than residues from A18 to F24, which are further in sequence from the labeled site. In the S29C* sample, residues from N12 to T15 displayed less broadening than residues from K5 to A11. The PRE effect for residues from T25 to D41 in the profiles of T17C and S29C displays an oscillating pattern, suggesting that helix or turn-like residual structures may be present in the region from T25 to D41, which is helical in the native state of protein L. This is consistent with the chemical shift perturbation in the region noted above. Comparison of the peaks in the experimental profile in Figure 7 row B to the profile expected for the native structure (Figure 7, row A) provides an indication of the extent to which the residual structure in the denatured state is native-like. The peaks in the region corresponding to the first β-hairpin (near residue 21 in E1C, residue 7 in S29 and T46) in the experimental profiles mirror peaks in similar locations in the profiles derived from the native structure (row A). In contrast, the peak in the region corresponding to the second β-hairpin in the C63 derivative (near residue 51) has no counterpart in the native profile. These results suggest that native-like chain reversals are sampled in the region corresponding to the first β-hairpin in the denatured state, but in the region corresponding to the second β-hairpin there is a chain reversal, not in the β-turn, but near the middle of the last strand. Comparison of the experimental profiles in row B to those for the simulated denatured state model with the local sequence-structure propensities of the protein L sequence (Figure 7, row C) provides some insight into the origin of the residual structure in the protein L denatured state. A peak observed in the experimental PRE profile in the vicinity of the first hairpin in the E1C derivative is also observed in the sequence-specific

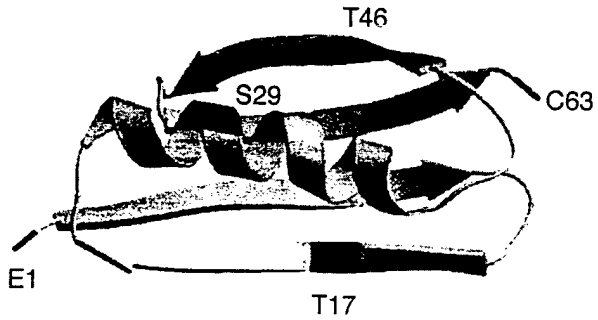


Figure 4. Ribbon diagram of the protein L structure. The positions where the spin labels were introduced are highlighted in black.

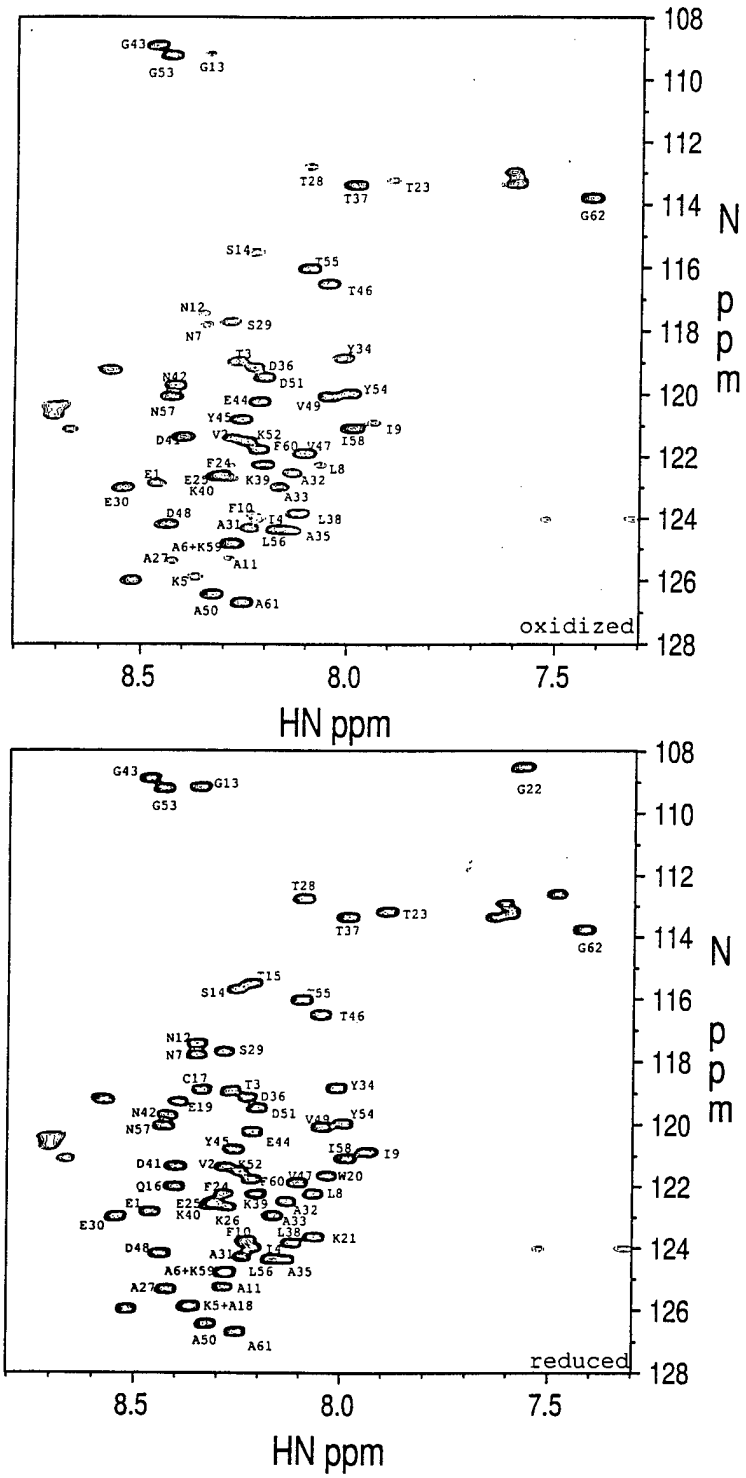


Figure 5. ^{1}H, ^{15}N-HSQC spectra of nitroxide-labeled T17C/F20W/Y32A mutant in 50 mM sodium phosphate and 2.2 M guanidine at pH 5.0 and 22 °C. (a) Oxidized form; (b) reduced form.

denatured state model simulations (row C), suggesting that the chain reversal in the vicinity of the first β-turn is due, at least in part, to local sequence propensities. Interestingly, the non-native peak observed around residue 51 in the experimental profile for the C63 derivative is also observed in the sequence-specific denatured state model, suggesting that this non-native feature is also due, in part, to local sequence propensities. However, much of the residual structure suggested by the experimental PRE profiles is likely to result from interactions longer in range than those captured by the denatured state model, as the oscillations more distant from the site of labeling in the experimental profiles in row B are mostly absent from the simulated profiles in row C.

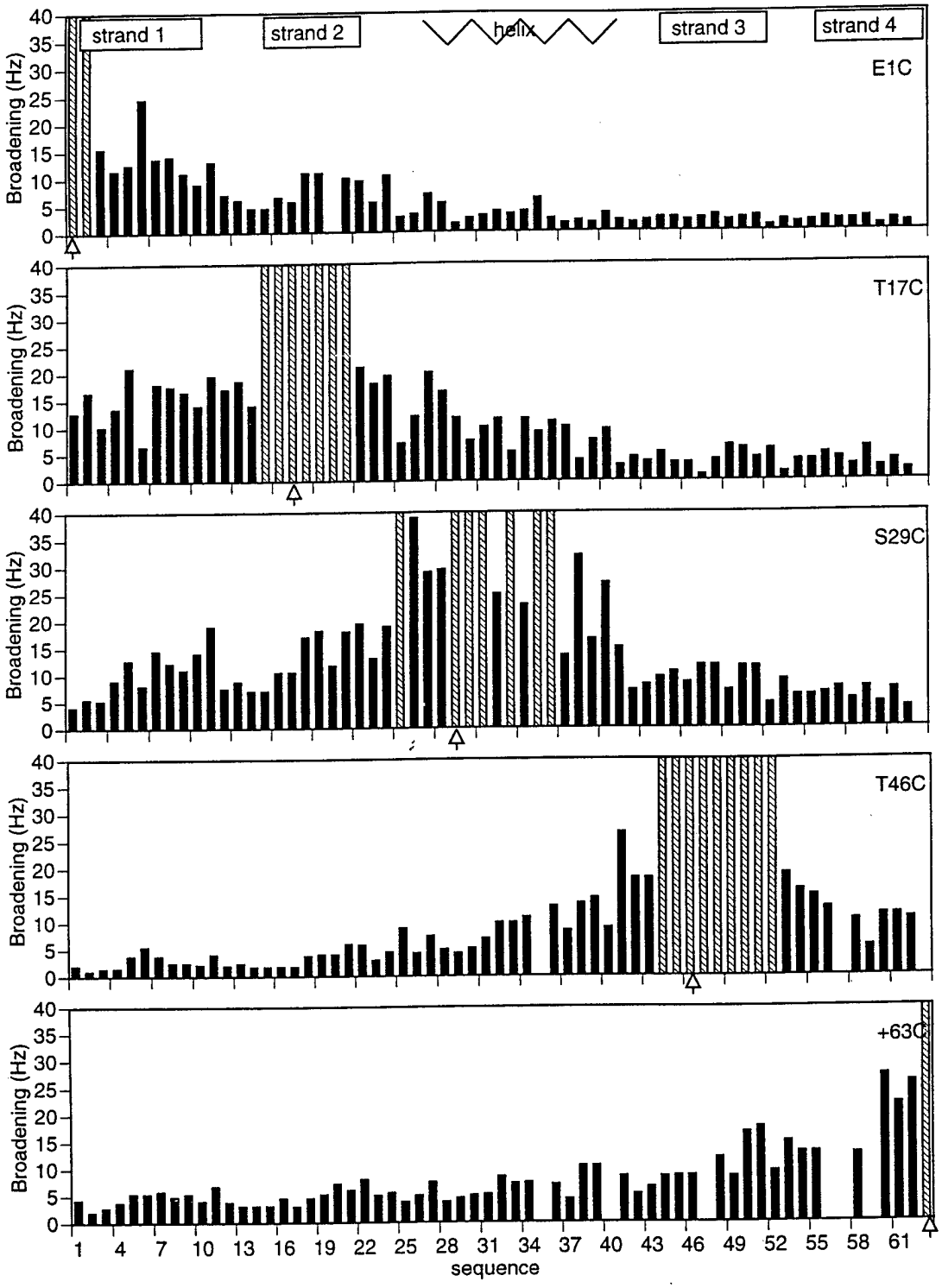


Figure 6. Paramagnetic relaxation enhancement of amide group proton resonances by the introduced nitroxide spin labels. The sites of labeling are indicated by the open triangles on the horizontal axes, and also labeled in the upper right corner of each profile. The hatched bars represent relaxation enhancement effects beyond the experimentally measurable limit. The secondary structure of native protein L is schematically represented on the top of the Figure.

Discussion

The NMR data presented here provide a picture of the conformations sampled in the denatured state of the F20W/Y32A mutant in 2 M guanidine. Almost all residues have chemical shifts close to their random coil values, and no long-range NOEs were observed even on perdeuterated samples, indicating considerable conformational averaging and little long-range order. The chemical shift data,

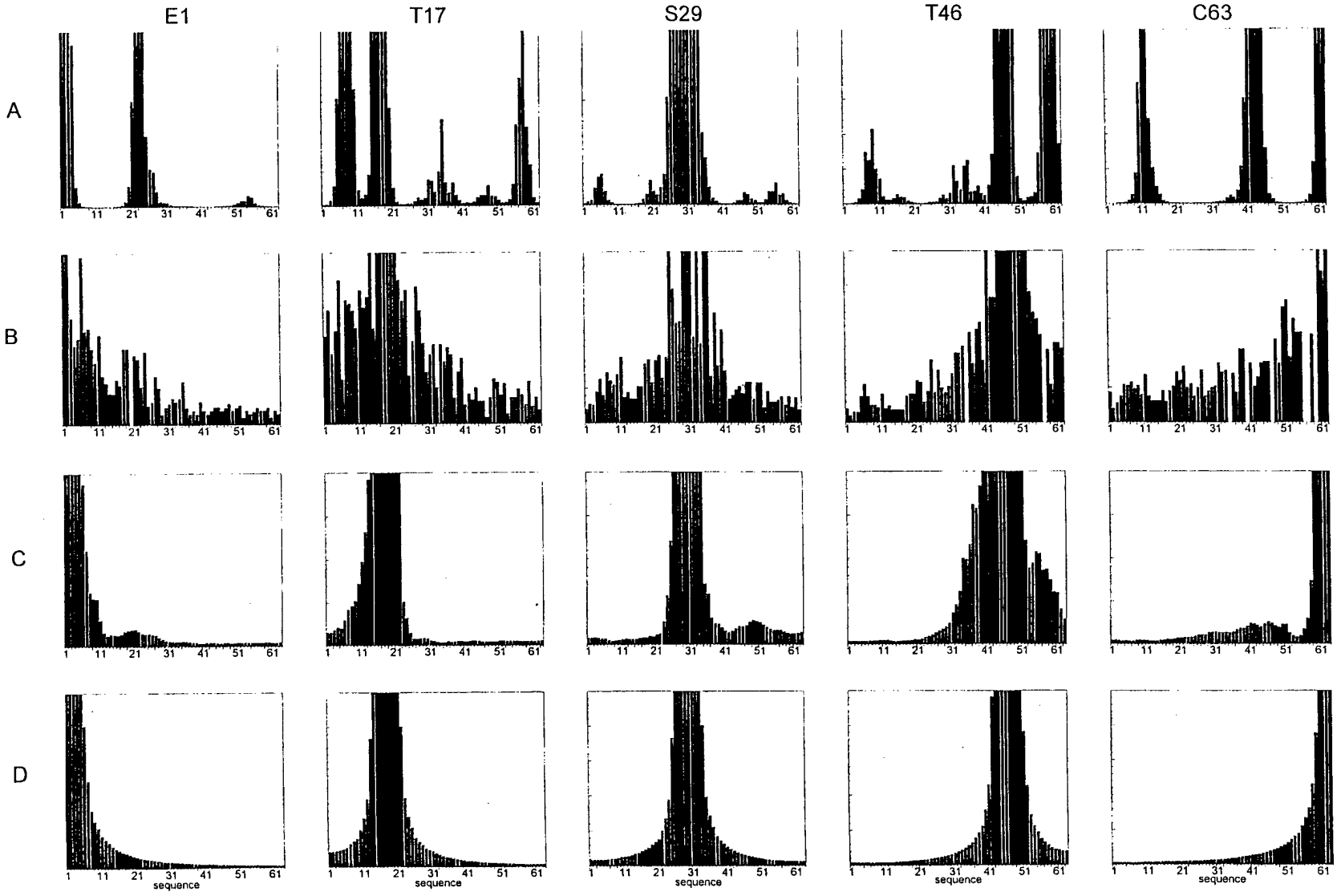


Figure 7. Comparison of experimental and simulated PRE profiles. Row A, simulated PRE profiles for the native protein L structure. Row B, experimental PRE profiles. Row C, simulated PRE profiles for denatured state model with protein L specific local sequence-structure biases. Row D, simulated PRE profiles for generic denatured state model without local sequence-structure biases. The simulation methods used in rows A, C, and D are described in Materials and Methods.

the observed medium-range NOEs, and the oscillations in the PRE profiles suggest that the greatest deviations from random chain behavior are in the N-terminal portion of the protein, and that these involve primarily native-like turns and chain reversals. The residues whose chemical shifts change the most with increasing guanidine concentration are just before and just after the central helix (G22, L38 and K39). The observed medium-range NOEs were in regions corresponding to the turn between the first two β-strands (I9 to G13), the turn between the second strand and the helix (A18 to E25), and the turn following the helix (L38 to G43). The most pronounced minima in the PRE profiles were observed near the location of these turns in the native state, with the largest effect in the first β-turn (the dip between N12 and T15 is clearly evident in both the E1C and the S29 labeled proteins (Figure 6)). The absence of medium-range NOEs suggests that the second β-turn is significantly less populated than the other turns in the denatured state. The second β-turn containing three consecutive residues with positive φ angles, only one of which is a glycine residue (Wikstrom *et al.*, 1994). The turn is thus likely to be under considerable strain in the native state and, perhaps as a consequence of this, the turn appears to be largely disrupted in the rate-limiting step in the unfolding of protein L (Gu *et al.*, 1997). The absence of detectable NOEs in the second β-turn in the denatured state could also be due to such strain, and formation of the turn during folding may be driven by long-range interactions not present in the denatured state.

The NMR studies described here are consistent with the suggestion from earlier circular dichronism (CD) and fluorescence studies of non-random residual structure in the denatured state in the region corresponding to the first hairpin and the helix (Scalley *et al.*, 1999). The denaturant dependence of the fluorescence of W20 was found to be quite different in a ten-residue peptide derived from the protein L sequence (centered on W20) from that in the denatured protein, suggesting some residual structure around W20 in the denatured protein. Consistent with this, we find that the highest density of medium-range NOEs in the denatured protein is around W20 (Figure 4). The earlier CD studies suggested some residual helix content in the denatured protein, and this is not inconsistent with the oscillations in the PRE profiles between T25 and D41 (this region is helical in the native state). Dead-time labeling HD exchange experiments on the denatured state immediately after initiation of refolding in the absence of denaturant suggest that the asymmetry observed in 2 M guanidine is also present in the absence of denaturant: the greatest protection from exchange was in the first β-hairpin.

A study of denatured proteins in urea and guanidine solutions also found little long-range order. A study of 434-repressor in 6 M urea (Neri *et al.*, 1992), revealed a native-like local hydrophobic cluster, with the remainder of the protein largely disordered. In a study of the urea and guanidine-denatured FK506 binding protein (Logan *et al.*, 1994), significant populations of both native-like and non-native-like residual local structures, mainly turns and local helical contents, were detected. More residual structure has been observed in denatured states of a truncated version of staphylococcal nuclease (Gillespie & Shortle, 1997) and the drk SH3 domain (Mok *et al.*, 1998) in the absence of denaturant. In the study of the unfolded drk SH3 domain, most of the observed long-range interactions disappeared upon the addition of 2 M guanidine, but there were still a few long-range NOEs detected in 2 M guanidine (Mok *et al.*, 1998). NMR characterization of the denatured states of drk and staphylococcal nuclease in the absence of denaturant was made possible by their relatively high level of solubility; unfortunately, we have not been able to identify unfolded protein L mutants which are sufficiently soluble in the absence of denaturant for NMR studies. The origins of the differences in solubility of the denatured states of different proteins are not entirely clear; the high level of solubility of the staphylococcal nuclease denatured state may, in part, be due to electrostatic repulsion between monomers resulting from the high net charge on the protein.

An issue of considerable current interest is how residual structure in the unfolded state contributes to the overall folding reaction. On the one hand, native-like interactions in the denatured states can limit the conformational search space during folding and, on the other, non-native like interactions could create energy barriers that hamper protein folding. For example, in the case of the drk SH3 domain, non-native interactions in the unfolded state destabilize the native state and may slow the folding process. Extensive characterization of the effects of mutations on protein L folding (Kim *et al.*, 1999) has suggested that the first β-turn is largely formed, and the second β-turn, largely disrupted in the folding transition state ensemble. Here, we observe medium-range NOEs and a significant dip in the PRE profile in the region corresponding to the first β-turn, but not the second, consistent with the asymmetry of structure in the folding transition state. Thus, the interactions formed in the protein L denatured state do not appear to disfavor subsequent folding events, and the asymmetry in protein L folding appears quite early in the folding process. The rate-limiting step in folding may involve the entropically costly consolidation and/or association of parts of the protein which are partially ordered in the denatured state.

Materials and Methods

Sample preparation

^{15}N-labeled and ^{15}N, ^{13}C-labeled protein samples were made by growing the transformed *Escherichia coli* cells in

Mops minimal medium. For the labeling, 99.9% (w/w) $^{15}NH_4Cl$ was used as the nitrogen source and 99.9% (w/w) ^{13}C-glucose as the carbon source in our media. For ^{15}N, ^{2}H-labeled F20W/Y32A, a single colony of *E. coli* of BL21 (DE3/plysS) carrying the F20W/Y32A plasmid was inoculated into 100 ml of M9 H_2O media with 50 μg/ml carbenicillin, and grown at 37 °C until the absorbance at 600 nm was approximately 0.6. The cells were harvested by centrifuging the culture with an ss-34 rotor at 4000 rpm for ten minutes. The harvested cells were resuspended with 10 ml of M9 $^{2}H_2O$ medium and then transferred into 2 l of M9 $^{2}H_2O$ medium with 50 μg/ml carbenicillin. The cells were grown at 37 °C to an absorbance at 600 nm of approximately 0.6, then induced with 1 mM IPTG for ten hours before harvesting. All labeled F20W/Y32A proteins were purified by His-tag affinity column as described (Gu *et al.*, 1995). For the ^{15}N, ^{2}H-labeled F20W/Y32A, the level of deuteration was estimated to greater than 95% based on the result from mass spectrometry.

Nitroxide labeling

Nitroxide groups can be introduced into proteins through alkylation of the thiolate group of cysteine residue. Since protein L does not contain any cysteine residues, site-directed mutagenesis was used to substitute selected amino acid residues with cysteine. Five charged or polar residues with highly solvent-exposed sidechains were selected for cysteine mutation to minimize possible conformational perturbations of mutagenesis. All the mutants, E1C, T17C, S29C, T46C and +63C (addition of a Cys residue to the C terminus), were spin labeled as described (Mchaourab *et al.*, 1996). The extent of labeling was examined by MALDI and in all cases it was greater than 95%.

NMR experiments and data processing

Both HNCACB (Wittekind & Mueller, 1993) and CBCACONH (Grzesiek & Bax, 1992, 1993) triple resonance experiments were carried out on a Bruker DMX500 instrument with 1.5 mM ^{15}N, ^{13}C-labeled F20W/Y32A in 10% $^{2}H_2O$/90% H_2O 50 mM sodium phosphate and 2.2 M guanidine at pH 5.0 and 22 °C. Matrices of 40 × 40 × 512 complex points were acquired with spectral widths of 7002.8, 2000.0, and 4496.4 Hz (F1, F2, F3) for both HNCACB and CBCACONH. For both ^{13}C and ^{15}N dimensions (F1, F2) of these spectra, the sizes of the time domain were doubled *via* forward-back linear prediction (Zhu & Bax, 1992). The data were zero-filled and extracted (only 4.70 ppm-9.00 ppm of the acquisition dimension was retained) to give final 3D data sets of 1024 × 80 × 80 real points. The guanidine titration was carried out using ^{15}N-labeled F20W/Y32A/C63 protein; the HSQC spectrum of this protein, which was readily purified in large amounts, is nearly identical with that of F20W/Y32A. HSQC experiments were carried out on samples equilibrated with 50 mM sodium phosphate and guanidine concentrations of 1.5 M, 2.0 M, 2.5 M, 3.0 M, 3.5 M, 4.0 M and 5.0 M at pH 5.0 and 22 °C.

The ^{1}H, ^{15}N-HSQC-NOESY-HSQC (Zhang *et al.*, 1997) experiment was performed on a three-channel Varian Inova 500 MHz spectrometer with a 1.2 mM perdeuterated ^{2}H, ^{15}N-labeled F20W/Y32A in 10% $^{2}H_2O$/90% H_2O 50 mM sodium phosphate and 2.2 M guanidine at pH 5.0 and 5 °C. The HSQC spectrum of F20W/Y32A under 2.2 M guanidine at pH 5.0 and 5 °C was almost identical with that under the same solvent conditions at 22 °C, suggesting that the population of unfolded F20W/Y32A was not changed significantly by varying the temperature between 22 °C and 5 °C. A matrix of 64 × 32 × 512 complex points was acquired with spectral widths of 1500.0, 1500.0 and 9000.9 Hz (F1, F2 and F3) using a mixing time of 600 ms and a recycle delay of 1.9 s. Some 24 scans were acquired for each FID. For both ^{15}N dimensions (F1 and F2), the sizes of the time domain were doubled *via* forward-backward linear prediction. The data were apodized with a 65°-shifted squared sine-bell in all three dimensions, zero-filled and extracted (only 4.70 ppm to 11.0 ppm was retained) to give a final 3D data set of 512 × 256 × 128 real points. All the NMR spectra were processed using the NMRPipe software system (Delaglio *et al.*, 1995).

To measure the paramagnetic relaxation enhancement due to the introduced nitroxide spin labels, ^{1}H, ^{15}N-HSQC spectra were collected using the pulse sequence described (Kay *et al.*, 1992) on 0.5 to 1.0 mM protein samples at pH 5.0, 22 °C before and after reduction by a threefold molar excess of ascorbic acid in a 5 μl volume. All the spectra were apodized with a 54°-shifted squared sine-bell in both dimensions and zero-filled. The intensities of peaks in the HSQC spectra of both the oxidized and the reduced forms were measured. The effects of paramagnetic enhancement were determined by spectral simulation as described below (the details are described by Gillespie & Shortle, 1997). First, the PRE effects on a set of Lorentzian peaks were simulated by multiplying their FIDs with an exponential window function using varying amounts of line broadening. This simulates the transverse relaxation of the amide group protons beginning with the first 90° pulse of HSQC experiment. To simulate relaxation during the pulse sequence of the HSQC experiment prior to signal acquisition, the first 18 ms of the FID was discarded because the HSQC pulse sequence used in this study involves a total of 18 ms of fixed delay at which the amide group proton magnetization resides in the transverse plane prior to signal acquisition. The remainder of the FID was Fourier-transformed after apodizing with a 54°-shifted squared sine-bell and zero-filled. Since the decrease in peak intensity depends on both the initial linewidths and the time constant of the exponential window function, two sets of simulation curves (basically, a plot of relative intensity *versus* broadening in Hz) were obtained for resonances with linewidths corresponding to those measured in F20W/Y32A, namely 15 and 20 Hz. Based on these simulation curves, the amount of line broadening corresponding to the experimentally measured intensity ratio of the oxidized *versus* the reduced forms was considered to be the paramagnetic relaxation enhancement (y-axis in Figure 6).

Simulation of the PRE effects

For the native state of protein L, PRE data were simulated directly from the protein structure according to the Soloman-Bloembergen equations. Interaction distances were measured between each pair of alpha-carbon and backbone nitrogen atoms in the wild-type structure, interaction strengths were taken to be proportional to the reciprocal of the sixth power of that distance, and distances closer than 6 Å were truncated to 6 Å. It should be noted that the actual position of the free electron is some distance from the alpha-carbon atom, and thus the simulated spectra are not expected to match the experimental spectra at very short sequence separations

(the lack of oscillation in the helix region in the simulated native spectrum, for example, is because all alpha-carbon-amide nitrogen pairs separated by less than four residues in the helix are separated by less than 6 Å). For simulations of disordered protein L, two different models were used to create representative ensembles of disordered structures, and the ensemble average signal between each pair of residues was calculated. First, an ensemble of structures built from unrelated protein fragments was used as a generic model of the states accessible to a disordered protein chain. Each structure was assembled by ligating three residue fragments picked at random from a 155,000-residue database of protein structures in which each entry had less than 40% sequence homology to all others. The geometry of each fragment in the assembled structure was determined by its ϕ, ψ, and ω angles in the database using a set of ideal bond angles and lengths (Engh *et al.*, 1991; the torsion angles were optimized to reproduce the native structures using the ideal bond lengths and angles (Simons *et al.*, 1997)). In the buildup procedure, side-chains were approximated with centroids. Structures with severe steric clashes were discarded. Second, to model a disordered chain with local interactions favored by the protein L sequence, an ensemble of structures was assembled as for the generic sequence model, but only three residue fragments (from the same database) with sequence identity to protein L were used. There were about 30-50 different fragments in the database with the correct sequence for each position in the protein. Local steric clashes were reduced by requiring that the four residues overlapping each junction between two fragments were represented by a four-residue fragment of similar sequence and structure in the database. Structures with severe steric clashes were discarded.

Acknowledgments

We thank Tanya Kortemme and David Shortle for helpful comments on the manuscript. This work was supported by a grant from the NIH and a young investigator award from the Packard foundation to D.B.

References

Braun, D., Wider, G. & Wuthrich, K. (1994). Sequence-corrected ^{15}N "random coil" chemical shifts. *J. Am. Chem. Soc.* **116**, 8466-8469.

Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. & Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*, **6**, 227-293.

Eliezer, D., Yao, J., Dyson, H. J. & Wright, P. E. (1998). Structural and dynamic characterization of partially folded states of apomyoglobin and implications for protein folding. *Nature Struct. Biol.* **5**, 148-155.

Engh, R. A. & Huber, R. (1991). Accurate bond and angle parameters for X-ray protein-structure refinement. *Acta Crystallog. sect. A*, **47**, 392-400.

Fong, S., Bycroft, M., Clarke, J. & Freund, M. V. (1998). Characterization of urea-denatured states of an immunoglobulin superfamily domain by heteronuclear NMR. *J. Mol. Biol.* **278**, 417-429.

Gillespie, J. & Shortle, D. (1997). Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels. *J. Mol. Biol.* **268**, 158-169.

Grzesiek, S. & Bax, A. (1992). Correlating backbone amide and side-chain resonances in larger proteins by multiple relayed triple resonance NMR. *J. Am. Chem. Soc.* **114**, 6291-6293.

Grzesiek, S. & Bax, A. (1993). Amino acid type determination in the sequential assignment procedure of uniformly ^{13}C/^{15}N enriched proteins. *J. Biomol. NMR*, **3**, 185-204.

Gu, H., Yi, Q., Bray, S. T., Riddle, D. S., Shiau, A. K. & Baker, D. (1995). A phage display system for studying the sequence determinants of protein folding. *Protein Sci.* **4**, 1108-1117.

Gu, H., Kim, D. & Baker, D. (1997). Contrasting roles for symmetrically disposed β-turns in the folding of a small protein. *J. Mol. Biol.* **274**, 588-596.

Kay, L. E., Keifer, P. & Saarinen, T. (1992). Pure absorpiton gradient enhanced heteronuclear single quantum correlation spectroscopy with improved sensitivity. *J. Am. Chem. Soc.* **114**, 10663-10665.

Kim, D. E., Fisher, C. & Baker, D. (2000). A breakdown of symmetry in the folding transition of protein L. *J. Mol. Biol.* In the press.

Kosen, P. A. (1989). Spin labeling of proteins. *Methods Enzymol.* **177**, 86-121.

Logan, T. M., Theriault, Y. & Fesik, S. W. (1994). Structural characterization of the FK 506 binding protein unfolded in the urea and guanidine hydrochloride. *J. Mol. Biol.* **236**, 637-648.

Mchaourab, H. S., Lietzow, M. A., Hideg, K. & Hubbell, W. L. (1996). Motion of spin-labeled side chains in T4 lysozyme. Correlation with protein structure and dynamics. *Biochemistry*, **35**, 7692-7704.

Mok, Y., Kay, C. M., Kay, L. E. & Forman-Kay, J. D. (1998). NOE data demonstrating a compact state for an SH3 domain under non-denaturing conditions. *J. Mol. Biol.* **289**, 619-638.

Neri, D., Billeter, M., Wider, G. & Wuthrich, K. (1992). NMR determination of residual structure in a urea-denatured protein, the 434-repressor. *Science*, **257**, 1559-1563.

Plaxco, K. W., Millett, I. S., Segel, D. J., Doniach, S. & Baker, D. (1999). Chain collapse can occur concomitantly with the rate-limiting step in protein folding. *Nature Struct. Biol.* **6**, 554-556.

Sattler, M. & Fesik, S. W. (1996). Use of deuterium labeling in the NMR: overcoming a sizable problem. *Structure*, **4**, 1245-1249.

Scalley, M. L., Nauli, S., Galdwin, S. T. & Baker, D. (1999). Structural transitions in the protein L denatured state ensemble. *Biochemistry*, **38**, 15297-15335.

Scalley, M. L., Yi, Q., Gu, H., McCormack, A., Yates, J. R., III & Baker, D. (1997). Kinetics of folding of the IgG binding domain of peptosptreptococcal protein L. *Biochemistry*, **36**, 3373-3382.

Shortle, D. (1996a). The denatured state (the other half of the folding equation) and its role in protein stability. *FASEB J.* **10**, 27-34.

Shortle, D. (1996b). Structural analysis of non-native states of proteins by NMR methods. *Curr. Opin. Struct. Biol.* **6**, 24-30.

Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209-225.

Solomon, I. & Bloembergen, N. (1956). Nuclear magnetic interactions in the HF molecule. *J. Chem. Phys.* **25**, 261-266.

Spera, S. & Bax, A. (1991). Empirical correlation between protein backbone conformation and Ca and Cb ^{13}C nuclear magnetic resonance chemical shifts. *J. Am. Chem. Soc.* **113**, 5490-5492.

Wikstrom, M., Drakenberg, T., Forsen, S., Sjobring, U. & Bjorck, L. (1994). Three-dimensional solution structure of an immunoglobulin light chain-binding domain of protein L. Comparison with the IgG-binding domains of protein G. *Biochemistry*, **33**, 14011-14017.

Wishart, D. S. & Sykes, B. D. (1994). The ^{13}C chemical shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical shift data. *J. Biomol. NMR*, **4**, 171-180.

Wittekind, M. & Muller, L. (1993). HNCACB, a high-sensitivity 3D NMR experiment to correlate amide-proton and nitrogen resonances with the alpha- and beta-carbon resonances in proteins. *J. Magn. Reson.* **101**, 201-205.

Yao, J., Dyson, H. J. & Wright, P. E. (1997). Chemical shift dispersion and secondary structure prediction in unfolded and partly-folded proteins. *FEBS Letters*, **419**, 285-289.

Yi, Q. & Baker, D. (1996). Direct evidence for a two-state protein unfolding transition from hydrogen-deuterium exchange mass spectrometry and NMR. *Protein Sci.* **5**, 1060-1066.

Yi., Q., Scalley, M. L., Simons, K. T., Gladwin, S. & Baker, D. (1997). Characterization of the free energy spectrum of peptostreptococcal protein L. *Fold. Des.* **2**, 271-280.

Zhang, O., Forman-Kay, J. D., Shortle, D. & Kay, L. E. (1997). Triple-resonance NOESY-based experiments with improved spectral resolution: applications to structural characterization of unfolded partially folded and folded proteins. *J. Biol. NMR*, **9**, 181-200.

Zhu, G. & Bax, A. (1992). Two-dimensional linear prediction for signals truncated in both dimensions. *J. Magn. Reson.* **98**, 192-199.

Edited by P. E. Wright

(Received 22 February 2000; received in revised form 18 April 2000; accepted 24 April 2000)

A "loop entropy reduction" phage-display selection for folded amino acid sequences

PHILIPPE MINARD,[1,3,4] MICHELLE SCALLEY-KIM,[2,3] ALEX WATTERS,[2,3] AND DAVID BAKER[1]

[1]Department of Biochemistry, Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA

[2]Molecular and Cellular Biology Program, University of Washington, Seattle, Washington 98195, USA

(RECEIVED August 1, 2000; ACCEPTED October 27, 2000)

Abstract

As a step toward selecting folded proteins from libraries of randomized sequences, we have designed a 'loop entropy reduction'-based phage-display method. The basic premise is that insertion of a long disordered sequence into a loop of a host protein will substantially destabilize the host because of the entropic cost of closing a loop in a disordered chain. If the inserted sequence spontaneously folds into a stable structure with the N and C termini close in space, however, this entropic cost is diminished. The host protein function can, therefore, be used to select folded inserted sequences without relying on specific properties of the inserted sequence. This principle is tested using the IgG binding domain of protein L and the lck SH2 domain as host proteins. The results indicate that the loop entropy reduction screen is capable of discriminating folded from unfolded sequences when the proper host protein and insertion point are chosen.

Keywords: Phage-display; insertion; protein folding; protein evolution; chimeric proteins

What fraction of the vast number of possible polypeptide sequences are able to form a defined three-dimensional structure analogous to the folded states of natural proteins? An experimental answer to this fundamental question could, in principle, be obtained by an examination of the properties of a large number of randomly generated polypeptide sequences. We have developed a 'loop entropy reduction' screen to be used for selecting folded proteins from large collections of natural or artificial coding sequences.

Recent studies have shown that increasing the length of loop regions in protein structures typically results in a decrease in overall stability of the protein. The observed decrease in stability has been linked to the entropic cost of ordering the additional residues in the loop (Ladurner 1997; Nagi 1997). These data suggest that if the unfolded inserted sequence is of sufficient length, it may disrupt the folding of the host protein into which it is inserted. However, if the inserted sequence is capable of folding into an independent structure, the entropic cost to the host protein will be minimal and the host protein may retain its ability to fold (Betton et al. 1997; Collinet 2000). The functional integrity of the host protein is directly related to the conformational state of the inserted sequence, allowing the properties of the host protein to be used to select for folded inserted sequences.

Several questions must be addressed in the design of the loop entropy reduction screen: What protein should serve as the host protein? Into which loop should the sequences be inserted? How will the folded state of the host protein be evaluated? In this study, we have chosen to use the IgG-binding domain of protein L and the lck SH2 domain as host proteins. These proteins were selected because of the large amount of structural information available: High-resolution structures and thermodynamic stabilities have been deter-

Reprint requests to: David Baker, Department of Biochemistry, University of Washington, J 567 Health Sciences Building, Box 37-7350, Seattle, Washington 98195, USA; e-mail: dabaker@u.washington.edu; fax: 206-685-1792.

[3]These authors contributed equally to this work.

[4]Present address: Laboratoire de Modelisation et Ingenierie des Proteines, CNRS et Universite de Paris sud, 91405 CEDEX ORSAY, France.

Article and publication are at www.proteinscience.org/cgi/doi/10.1110/ps.32401

mined for both proteins (Wikstrom et al. 1993; Tong et al. 1996; Scalley et al. 1997). A phage-display format was chosen as a screening method because it has proven to be successful in selecting rare folded variants within a collection of highly randomized domains (Zhou et al. 1996; Riddle et al. 1997; Kim et al. 1998). Furthermore, the resistance of filamentous phage is compatible with the presence of destabilizing conditions such as high temperature or denaturants during the selection procedure, providing a method for adjusting the level of selection pressure (Kristensen 1998; Forrer et al. 1999; Jung et al. 1999).

Results

Initially, the IgG-binding domain of peptostreptococcal protein L was selected as a host protein. Two turns were chosen as points for insertion (Figure 1): the turn leading from the second β-strand into the α-helix, β_2–α, and the turn leading from the third β-strand into the fourth β-strand, β_3–β_4. To determine whether either or both insertion points could withstand the insertion of a folded sequence, preliminary experiments were performed using wild-type src SH3 as a model of a folded inserted sequence. The SH3 domain coding region was introduced into the protein L DNA sequence via two unique restriction sites (*Eco*RI/*Nde*I for β_2–α, and *Sal*I/*Kpn*I for β_3–β_4) in the protein L-gene VIII fusion construct (Gu et al. 1995). The amino acid sequences of the regions bordering the SH3 insertion are shown in Table 1. It is important to note that N and C termini of SH3 domains are close in space to one another, a property that is likely to be preferred by the screening method.

Phage displaying the chimeric proteins were made and screened for their ability to bind paramagnetic beads coated with IgG, which protein L binds to with high affinity (Kihlberg et al. 1996). The results show clearly that the turns cannot tolerate the insertion of SH3 equally (Table 2). Phage displaying SH3 inserted into the β_2–α turn were recovered at very low levels, whereas phage displaying SH3 inserted into the β_3–β_4 turn of protein L were recovered at levels similar to that of phage displaying wild-type protein L. Recently it was found that the residues packed at the β_2–α interface make important contributions to IgG binding, thus, the low recovery levels may reflect a disruption in IgG binding resulting from the proximity of the SH3 insertion (H. Svensson, pers. comm.). Because the β_3–β_4 turn tolerates insertion of a folded protein, this turn was used for all subsequent insertions into protein L.

A library of random sequences coding for 60–300 amino acids was constructed and inserted into the β_3–β_4 turn of protein L (see Materials and Methods). The library members that retained protein L function were isolated using IgG-coated beads. Several positives were identified by the screen, but initial examination of their sequences showed

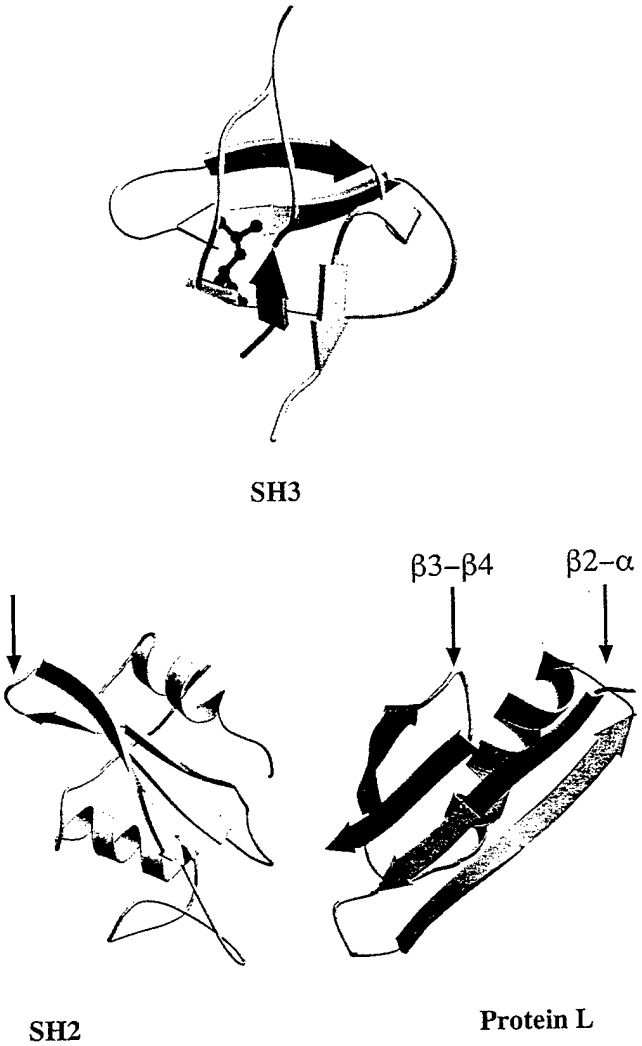


Fig. 1. Structure of the host and inserted domains. The surface loops in protein L and SH2 that are used for insertion are indicated by the arrows. The position of the A8G mutation in protein L is highlighted on the structure. Leu32 is mutated to a glutamate in the destabilized form of the SH3 domain. Images were made using Raster 3D (Bacon and Anderson 1988; Merrit and Murphy 1994) and Molscript (Kraulis 1991).

them to be highly polar and, thus, unlikely to fold into stable independent structures. One sequence identified by the screen, B11, was chosen for further analysis (see the legend to Table 1 for B11 sequence). Phage displaying the pL[B11] chimera were panned against IgG-coated beads to ensure the chimera exhibited binding activity (Table 2). The pL[B11] chimera was recovered at levels similar to that of wild-type protein L and the pL[SH3] chimera.

To examine the effect of folded and unfolded inserted sequences on the stability of protein L, the SH3 and B11 chimeric proteins, pL[SH3] and pL[B11], were overexpressed and purified. The stabilities of pL[SH3] and pL[B11] were measured using guanidine denaturation (Fig. 2; see Materials and Methods). Both pL[SH3] and pL[B11]

Table 1. *Phagemid constructs*

Host	Insert	Sequence
$pL_{\beta 2-\alpha}$	—	$T_{19}A_{20}[E_{21}F_{22}K_{23}G_{24}T_{25}F_{26}E_{27}K_{28}A_{29}T_{30}S_{31}E_{32}A_{33}Y_{34}A_{35}]Y_{36}A_{37}$
$pL_{\beta 2-\alpha}$	$SH3_{wt}$	$T_{19}A_{20}[E_{21}F_{22}K_{23}G_{24}T_{25}F_{26}GT_9F_{10}A_{11}L_{12}...(SH3)...A_{62}P_{63}S_{64}D_{65}GE_{27}K_{28}A_{29}T_{30}S_{31}E_{32}A_{33}Y_{34}A_{35}]Y_{36}A_{37}$
$pL_{\beta 3-\beta 4}$	—	$T_{48}V_{49}[D_{50}V_{51}A_{52}D_{53}K_{54}G_{35}]T^*Y_{56}T_{57}L_{58}$
$pL_{\beta 3-\beta 4}$	$SH3_{wt}$	$T_{48}V_{49}[E^*_{50}V_{51}A_{52}D_{53}K_{54}SGT_9F_{10}A_{11}L_{12}...(SH3)...A_{62}P_{63}S_{64}D_{65}G_{55}]T^*Y_{56}T_{57}L_{58}$
$pL_{\beta 3-\beta 4}$	B11	$T_{48}V_{49}[E^*_{50}V_{51}A_{52}D_{53}K_{54}SGR_1S_2P_3A_4...(B11)...I_{58}D_{59}T_{60}G_{61}S_{62}G_{55}]T^*Y_{56}T_{57}L_{58}$
SH2	—	$R_{168}D_{169}V^*_{170}[D_{171}Q_{172}N_{173}Q_{174}G_{175}]T^*_{176}V_{177}V_{178}$
SH2	$SH3_{wt}$	$R_{168}D_{169}V^*_{170}[EVADKSGT_9F_{10}A_{11}L_{12}...(SH3)...A_{62}P_{63}S_{64}D_{65}G_{175}]T^*_{176}V_{177}V_{178}$
SH2	B11	$R_{168}D_{169}V^*_{170}[EVADKSGR_1S_2P_3A_4...(B11)...I_{58}D_{59}T_{60}G_{61}S_{62}G_{175}]T^*_{176}V_{177}V_{178}$
SH2	$SH3_{L32E}$	$R_{168}D_{169}V^*_{170}[EVADKSGT_9F_{10}A_{11}L_{12}...(SH3)...E^*_{32}...A_{62}P_{63}S_{64}D_{65}G_{175}]T^*_{176}V_{177}V_{178}$

The sequence of the inserted protein and the linker region are typed in bold and italics, respectively. The position corresponding to the restriction site (EcoRI/NdeI) for $pL_{\beta 2-\alpha}$, SalI/KpnI for SH2 and $pL_{\beta 3-\beta 4\ a}$) are in brackets. Introduction of the KpnI site into protein L results in the insertion of a threonine residue between G_{55} and Y_{56}. Two point mutations, F170V and E176T, were introduced into SH2 to create the SalI and KpnI restriction sites and are indicated with *. Control experiments have shown that these mutations do not impact the binding activity of phage displaying protein L of SH2. The sequence of B11 follows: RSPAQVVDAQQNAVKDNEPSGSALGGRSAPGATRPSDQSGGSEDRSVPTEKPKEGPHID. Sequence numbering systems for SH2 and SH3 are described in Tong et al. (1996) and Riddle et al. (1997), respectively.

exhibited a cooperative and reversible folding transition with *m*-values (the denaturant dependence of the free energy of folding) similar to that of wild-type protein L (wild-type protein L: $m = 1.8$; pL[SH3]: $m = 1.7$; pL[B11]: $m = 1.8$). The free energy of unfolding for pL[SH3] was slightly reduced in comparison with wild-type protein L, whereas that of pL[B11] was drastically decreased (wild-type protein L: $\Delta G = 4.6$ kcal/mol; pL[SH3]: $\Delta G = 3.4$ kcal/mol; pL[B11]: $\Delta G = 1.0$ kcal/mol). These data confirm our assumption that the insertion of an unfolded sequence into protein L results in a large decrease in stability, whereas the insertion of a folded sequence results in minimal stability loss.

The equilibrium denaturation data suggest that if protein L were destabilized by 1–2 kcal/mol, the folding of the pL[B11] chimera, but not that of the pL[SH3] chimera, would be disrupted, thereby decreasing the permissiveness of protein L with respect to loop insertions. To destabilize protein L, panning experiments with phage displaying wild-type protein L, pL[SH3], and pL[B11] were performed in 1 M guanidine (Table 2). In comparison with panning in the absence of guanidine, an approximate 10-fold loss in recovery was observed for wild-type protein L and pL[SH3] phage whereas a 100-fold loss in recovery was observed for pL[B11] phage. However, the recovery of pL[B11] was still 10-fold above background recovery levels.

Table 2. *Panning recoveries for protein L experiments*

Host	Insert	% Recovery 0M gnd	% Recovery 1M gnd
pL	none	1.0×10^{-2}	2.4×10^{-3}
$pL_{\beta 2-\alpha}$	SH3	6.0×10^{-4}	—
$pL_{\beta 3-\beta 4}$	SH3	2.0×10^{-1}	1.6×10^{-2}
pL	B11	5.0×10^{-2}	3.6×10^{-4}
pL_{A8G}	none	9.0×10^{-3}	1.2×10^{-3}
pL_{A8G}	SH3	5.0×10^{-3}	8.4×10^{-5}
pL_{A8G}	B11	8.2×10^{-4}	0
Control	n/a	1.2×10^{-5}	1.2×10^{-5}

5×10^9 c.f.u. of freshly prepared phage were used as input. % recovery is calculated as 100 × (c.f.u. input/c.f.u. output). Control experiments correspond to phage displaying the SH2 domain rather than protein L.

We then destabilized protein L by mutating residue A8 to a glycine (Fig. 1). This mutation was chosen because it destabilizes the protein by 2.4 kcal/mol and preliminary studies suggest that it does not interfere with IgG binding (Kim et al. 2000). This strategy of destabilizing the host by mutagenesis was used successfully in phage-display experiments using protein G, a small single-domain protein with a topology identical to protein L: A wild-type protein G host tolerated 50% of randomized turn sequences whereas a destabilized protein G host tolerated only a small fraction of the randomized sequences (Zhou et al. 1996). Both B11 and SH3 were inserted into the β_3–β_4 turn of the A8G point mutant and subjected to phage panning experiments. The results were similar to those observed for the wild-type protein L chimera panned in the presence of guanidine; compared with wild-type protein L, pL_{A8G} and pL_{A8G}[SH3] phage experienced a 10-fold loss in recovery and pL_{A8G} [B11] phage experienced a 100-fold loss in recovery. Because the recovery levels of pL_{A8G} [B11] phage were still 10-fold above background levels, additional panning experiments were performed in the presence of 1 M guanidine (Table 2). Both pL_{A8G}[SH3] and pL_{A8G}[B11] were recovered at very low levels, however, indicating that these conditions are too stringent for the pL_{A8G} chimeric proteins.

In an effort to find a less permissive host protein, we turned to the lck SH2 domain. A surface loop in SH2 (Fig. 1) was chosen as the insertion point because of its central location in the molecule and its lack of involvement in ligand binding. To test the efficacy of SH2 as a host protein, wild-type SH3, B11, and, as an additional model for an unfolded sequence, a strongly destabilized SH3 point mu-

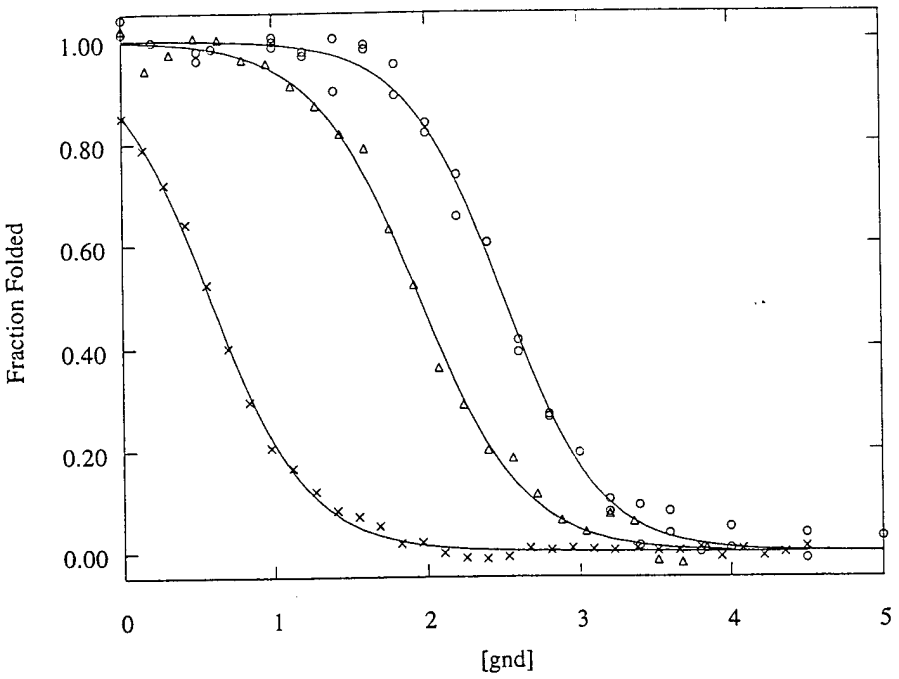


Fig. 2. Equilibrium denaturation melts of wild-type protein L, pL[SH3], and pL[B11]. Denaturation curves of wild-type protein L (open circles), pL[SH3] 0.(open triangles), and pL[B11] (crosses) were monitored by circular dichroism at 220 nm. The data were fit as described by Scalley et al. (1997). Protein concentrations were 10 ± 0.2 μM for wild-type protein L, 8 μM ± 0.2 μM for pL[B11], and 5 μM ± 0.2 μM for pL[SH3].

tant, $SH3_{L32E}$, were each inserted into the SH2 domain. Insertions were introduced into SH2 in a manner similar to protein L, using two unique restriction sites (*SalI*/*KpnI*) introduced into the SH2-gene VIII fusion construct. The amino acid sequences of the region bordering the insertions of wild-type SH3, $SH3_{L32E}$, and B11 into the SH2 domain are shown in Table 1.

To investigate the effects of the different insertions on SH2 function, phage displaying the SH2 insertion variants were panned using paramagnetic beads coated with SH2 ligand, a phosphotyrosyl peptide. The recovery of the different phage is reported in Table 3. The results show that the chimeric protein into which SH3 sequence was inserted is recovered with an efficiency on the same order of magnitude as the host protein without any insertion. The recovery is reduced to background levels when either $SH3_{L32E}$ or B11 is inserted into the host protein. These results indicate that the selection of phage displaying chimeric functional SH2 proteins can efficiently discriminate between folded and unfolded sequences without relying on any biological function of the inserted sequence.

Table 3. *Panning recoveries for SH2 experiments*

Host	Insert	% Recovery
SH2	none	1.0×10^{-2}
SH2	SH3	3.0×10^{-2}
SH2	SH_{L32E}	1.2×10^{-5}
SH2	B11	0
Control	n/a	2.4×10^{-5}

5×10^9 c.f.u. of freshly prepared phage were used as input. Control experiments correspond to phage displaying protein L rather than the SH2 domain.

Discussion

The results of this study show that a screen based on loop entropy reduction is capable of discriminating between folded and unfolded sequences. We have also shown that choosing the correct host protein and the position of insertion are critical to the success of the screen. For example, protein L will allow the insertion of a folded SH3 domain in the β_3–β_4 turn but not the β_3–α turn. However, the β_3–β_4 turn also accommodates the insertion of unfolded sequences, as seen by the retention of protein L function in the pL[B11] chimera.

By destabilizing protein L, we were able to reduce the recovery of the unfolded insert [B11] in comparison with the folded insert [SH3], although its recovery could not be reduced to background levels. The close proximity of the β_3–β_4 insertion site to the C terminus may contribute to the permissiveness of protein L because only a few residues are needed to complete the native structure, and a suitable replacement sequence may be found in the randomized insert.

In such a scenario, the remainder of the loop and the original C terminus would be extruded into the linker between protein L and gene VIII.

The underlying idea behind the loop entropy reduction screen is theoretically applicable to any host protein if a phenotypic screen related to the functional integrity of the host protein is available. A similar approach was employed using *Escherichia coli* RNase H1 as a host protein (Doi et al. 1997, 1998). In this study, random sequences were inserted into a surface loop of RNase H1, and chimeric proteins that retained RNase H1 function were selected using an in vivo assay. It was found that the inserted sequences that were folded maintained their structure on excision from the chimera. However, structural characterization of the chimeric proteins that came through the screen demonstrated that the inserted sequences were not always folded; three out of five of the chimeric proteins characterized were found to have unfolded inserted sequences. We observed a similar permissiveness with protein L as the host protein, providing further evidence that not all host proteins are equally good at discriminating between folded and unfolded inserts.

Other studies that have probed random sequence libraries for folded proteins have relied on expression of the random sequence as a screening method. Davidson and coworkers (Davidson and Sauer 1994; Davidson et al. 1995) constructed a library of random sequences consisting of three amino acids (Q, L, R). Interestingly, the fraction of sequences containing some degree of structure was large enough to allow detection of folded proteins using a screening method based on expression and solubility of protein from individual clones. In an extension of this work, Prijambada and coworkers (1996) constructed random sequence libraries containing all 20 amino acids. In that study, the investigators concluded that 8% of the random sequences were expressed and soluble, but no proteins with extensive secondary structure were found. To extend this work further, it is necessary to examine a larger number of sequences than is possible with expression-based screening methods.

In this study, we have employed a phage-display screening method that has several advantages over expression-based screening methods. First, phage-display techniques allows examination of many more sequences than is possible in an expression-based screen. Second, phage display has been proven successful in selecting rare folded variants within a collection of highly randomized domains (Zhou et al. 1996; Riddle et al. 1997; Kim et al. 1998). Additionally, the physical conditions of the selection step can be easily controlled and adapted to specific requirements. For instance, phage-panning experiments are compatible with the presence of reducing agents, denaturants, and proteases and these reagents effectively increase the selection pressure of the screen (Kristensen 1998; Jung et al. 1999). This feature may prove useful in eliminating false positives associated with inserted sequences that are marginally stable.

We are currently using the loop entropy reduction selection described in this paper to search for folded proteins in libraries of randomized synthetic sequences and libraries of shuffled genomic sequences. As a consequence of the design of the screen, it is very likely that only modules with their N and C termini in close proximity will be selected. Although this is certainly a limitation of the selection, it should be noted that the N and C termini are near one another in a disproportionately large number of globular proteins (Thornton 1983). It is attractive to speculate that this relatively high frequency of proximity between N and C termini is an evolutionary relic of a mechanism for generating complex, multidomain proteins from smaller folded units similar to our experimental selection strategy: the insertion of folded modules into loops of other folded modules. Thus, it is possible that the selection experiments may to some extent recapitulate the generation of the complex multidomain protein structures found in nature.

Materials and methods

Preparation and panning of phage

All phage were prepared as described in Gu et al. (1995). The preparation of IgG-coated magnetic beads and the subsequent protein L-panning experiments were performed as described in Gu et al. (1995), except for the guanidine-panning experiments where 1 M guanidine (USB, Ultrapure) was present in both the binding and washing steps. For the SH2-panning experiments, the streptavidin-coated magnetic beads (Dynabeads M-280) were coated with a biotinylated phosphotyrosine peptide (GGGGGGEPPQ[pY]EE IPIYL; synthesized by Sigma Genosys). A total of 20 μL of the streptavidin-coated beads (10 mg/mL) were incubated with 0.2 μg peptide, 0.5% TWEEN-TBS for 1 h, washed twice with 800 μL of 0.5% TWEEN-TBS, and resuspended in 20 μL of 0.5% TWEEN-TBS. The prepared beads (2 μL) were incubated with 5×10^9 phage particles for 1 h in 100 μL of a 4% milk, 0.1% TWEEN-TBS solution. The beads were washed 7 times with 800 μL of 0.5% TWEEN-TBS and resuspended in a final volume of 30 μL of 0.5% TWEEN-TBS. The phage bound to the beads were transfected into XLI Blue cells and plated onto LB agar with carbenicillin to quantitate the number of phage bound to the beads.

Random sequence library

A random sequence library was made by self-ligation of a highly degenerate cassette: GGATCC$(VNNNNB)_9$GGATC where N = A,C,T,G; V = C,A,G; B = G,T,C; and GGATCC is the *Bam*HI cleavage site. The VNN, NNB polymer codes for polypeptides containing all amino acids in proportions similar to the sequence of typical soluble proteins. The cassette is symmetrical, allowing polymerization in both orientations. A Gly–Ser sequence corresponding to the *Bam*HI sequence occurs every 20 amino acids in the polymer. As only one stop codon occurs among 96 possibilities, a significant fraction of polymerized cassettes inserted in the host protein sequences are expected to be full length. The cloning of the random sequences into the host protein sequence

requires the proper restriction sites and flanking sequences at each of the ends. Two adaptor cassettes ("Start" and "Stop") containing a single *Bam*HI-cohesive extremity and one uncleaved (*Sal*I and *Kpn*I, respectively, for start and stop) restriction site at the other end were introduced in a low molar ratio (1/8) in the random-cassette ligation reaction. Under these conditions, a ladder of products between 100 bp and 1 kb was obtained. For ligation into the host protein, the polymerized cassettes were cleaved with *Sal*I and *Kpn*I. Fragments >200 bp were gel purified and ligated into the vector. The constructs were then electroporated into XLI Blue *E. coli* cells to give a library of 1.5×10^5 independent clones. One round of phage-display selection followed by a colony-lift assay was sufficient to identify several positive clones. Two sequences containing 60 and 80 amino acids were sequenced and both displayed highly polar sequences. One of these sequences, B11, was used in subsequent experiments.

Equilibrium denaturation

The methods described by Gu et al. (1995) were used for the overexpression and purification of the chimeric proteins. Circular dichroism equilibrium denaturation experiments were performed as described by Scalley et al. (1997).

Acknowledgments

We thank members of the Baker laboratory for helpful discussion and especially Karen Butner and David Kim for their help with design and experimental procedures. This work was supported in part by a NATO Fellowship to P.M., NIH training grants to M.S.-K. and A.W., and a NIH grant to D.B.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Bacon, D.J. and Anderson, W.F. 1988. A fast algorithm for rendering space-filling molecule pictures. *J. Mol. Graph.* **6:** 219–220.

Betton, J.M., Jacob, J.P., Hofnung, M., and Broome-Smith, J.K. 1997. Creating a bifunctional protein by insertion of beta-lactamase into the maltodextrin-binding protein. *Nat. Biotechnol.* **15:** 1276–1279.

Collinet, B., Herve, M., Pecorari, F., Minard. P., Eder, O., and Desmadril, M. 2000. Functionally accepted insertions of proteins within protein domains. *J. Biol. Chem.* **275:** 17428–17433.

Davidson, A.R and Sauer, R.T. 1994. Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci.* **91:** 2146–2150.

Davidson, A.R., Lumb, K.J., and Sauer, R.T. 1995. Cooperatively folded proteins in random sequence libraries. *Nat. Struct. Biol.* **2:** 856–864.

Doi, N., Itaya, M., Yomo, T., Tokura, S., and Yanagawa, H. 1997. Insertion of foreign random sequences of 120 amino acid residues into an active enzyme. *FEBS Lett.* **402:** 177–180.

Doi, N., Yomo, T., Itaya, M., and Yanagawa, H. 1998. Characterization of random-sequence proteins displayed on the surface of Escherichia coli RNase HI. *FEBS Lett.* **427:** 51–54.

Forrer, P., Jung, S., and Pluckthun, A. 1999. Beyond binding: using phage display to select for structure, folding and enzymatic activity in proteins. *Curr. Opin. Struct. Biol.* **9:** 514–520.

Gu, H., Yi, Q., Bray, S.T., Riddle, D.S., Shiau, A.K., and Baker, D. 1995. A phage display system for studying the sequence determinants of protein folding. *Protein Sci.* **4:** 1108–1117.

Jung, S., Honegger, A., and Pluckthun, A. 1999. Selection for improved protein stability by phage display. *J. Mol. Biol.* **294:** 163–180.

Kihlberg, B.M., Sjoholm, A.G., Bjorck, L., and Sjobring, U. 1996. Characterization of the binding properties of protein LG, an immunoglobulin-binding hybrid protein. *Eur. J. Biochem.* **240:** 556–563.

Kim, D.E., Gu, H., and Baker, D. 1998. The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc. Natl. Acad. Sci.* **95:** 4982–4986.

Kim, D.E., Fisher, C., and Baker, D. 2000. A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* **298:** 971–984.

Kraulis, P.J. 1991. MOLSCRIPT- a program to produce both detailed and schematic plots of protei structures. *J. Appl. Crystallogr.* **24:** 946–950.

Kristensen, P.W. and Winter, G. 1998. Proteolytic selection for protein folding using filamentous bacteriophages. *Fold Des* **3:** 321–328.

Ladurner, A.G. and Fersht, A.R. 1997. Glutamine, alanine or glycine repeats inserted into the loop of a protein have minimal effects on stability and folding rates. *J. Mol. Biol.* **273:** 330–337.

Merrit, E.A. and Murphy, M.E.P. 1994. Raster 3D version 2.0. A program for photorealistic molecular graphics. *Acta Cyrstallogr.* sect. D 50, 869-873.

Nagi, A.D. and Regan, L. 1997. An inverse correlation between loop length and stability in a four-helix-bundle protein. *Fold Des* **2:** 67–75.

Prijambada, I.R., Yoma, T., Tanaka, F., Kawama, T., Yamamoto, K., Hasegawa, A., Shima, Y., Negora, S., and Urabe, I. 1996. Solubility of artificial proteins with random sequences. *FEBS Lett.* **382:** 21–25.

Riddle, D.S., Santiago, J.V., Bray-Hall, S.T., Doshi, N., Grantcharova, V.P., Yi, Q., and Baker, D. 1997. Functional rapidly folding proteins from simplified amino acid sequences. *Nat. Struct. Biol.* **4:** 805–809.

Scalley, M.L., Yi, Q., Gu, H., McCormack, A., Yates, J.R., and Baker, D. 1997. Kinetics of folding of the IgG binding domain of peptostreptococcal protein L. *Biochemistry* **36:** 3373–3382.

Thornton, J. and Sibanda, B. 1983. Amino and carboxy-terminal regions in globular proteins. *J. Mol. Biol.* **167:** 443–460.

Tong, L., Warren, T.C., King, J., Betageri, R., Rose, J., and Jakes, S. 1996. Crystal structures of the human p56lck SH2 domain in complex with two short phosphotyrosyl peptides at 1.0 A and 1.8 A resolution. *J. Mol. Biol.* **256:** 601–610.

Wikstrom, M., Sjobring, U., Kastern, W., Bjorck, L., Drakenberg, T., and Forsen, S. 1993. Proton nuclear magnetic resonance sequential assignments and secondary structure of an immunoglobulin lightchain-binding domain of protein L. *Biochemistry* **32:** 3381–3386.

Zhou, H.X., Hoess, R.H., and DeGrado, W.F. 1996. In vitro evolution of thermodynamically stable turns. *Nat. Struct. Biol.* **3:** 446–451.

doi:10.1006/jmbi.2000.4352 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) 306, 555–563

JMB

Circularization Changes the Folding Transition State of the src SH3 Domain

Viara P. Grantcharova and David Baker*

Department of Biochemistry
University of Washington
Seattle, WA 98195, USA

Native state topology has been implicated as a major determinant of protein-folding mechanisms. Here, we test experimentally the robustness of the src SH3-domain folding transition state to changes in topology by covalently constraining regions of the protein with disulfide crosslinks and then performing kinetic analysis on point mutations in the context of these modified proteins. Circularization (crosslinking the N and C termini) of the src SH3 domain makes the protein topologically symmetric and causes delocalization of structure in the transition state ensemble suggesting a change in the folding mechanism. In contrast, crosslinking a single structural element (the distal β-hairpin) which is an essential part of the transition state, results in a protein that folds 30 times faster, but does not change the distribution of structure in the transition state. As the transition states of distantly related SH3 domains were previously found to be very similar, we conclude that the free energy landscape of this protein family contains deep features which are relatively insensitive to sequence variations but can be altered by changes in topology.

Keywords: protein folding; folding kinetics; folding mechanism; transition state; SH3 domain

*Corresponding author

Introduction

A recent development in the protein folding field has been the empirical observation that native state topology is a major determinant of folding rates, with simple fold proteins folding faster than proteins with complicated topologies.[1] The remarkable correlation found between the average sequence separation of interacting residues in the native structure (contact order) and the rate of folding suggests that the free energy barrier to folding has a large entropic contribution while variations in the strength of the stabilizing interactions manifest themselves on a smaller scale. Consistent with the idea that the molecular details of the interactions are overshadowed by the entropic cost of making them, several theoretical models have been successful in predicting the transition state for folding and/or the folding rate for small proteins using only information from the native state structure.[2–5] Furthermore, recent experimental studies have established the conservation of folding transition states among homologous proteins with the same topology but sequence identity as low as 13%[6–8] suggesting that once the native-state topology is specified by the sequence the folding transition state is largely determined as well. The goal of our study is to explore and test these conclusions further.

Previous studies of the folding transition state of the src SH3 domain showed that it involves the association of the distal β-hairpin and the diverging turn, while the N and C termini are completely disordered (Figure 1(a)).[8,9] Here, we explore the robustness of this transition state to chain crosslinks in order to test the role of topology. Previously Serrano and co-workers showed that circularly permuting the α-spectrin SH3 domain can change its transition state depending on the site of permutation,[10] while smaller mutations that stabilize a part of the folding nucleus do not alter structure elsewhere in the transition state.[7,11] This argued for a conformationally restricted transition state, which requires the interaction of specific

Present address: V. P. Grantcharova, Center for Genomics Research, Harvard University, 16 Divinity Ave, Cambridge, MA 02138, USA.

Abbreviations used: SH3, src homology 3; Gnd, guanidine; wt, wild-type; CO, contact order; NC, circular src SH3 domain; SS, src SH3 mutant with a distal hairpin crosslink; CI2, chymotrypsin inhibitor 2.

E-mail address of the corresponding author: dabaker@u.washington.edu

0022-2836/01/030555–9 $35.00/0

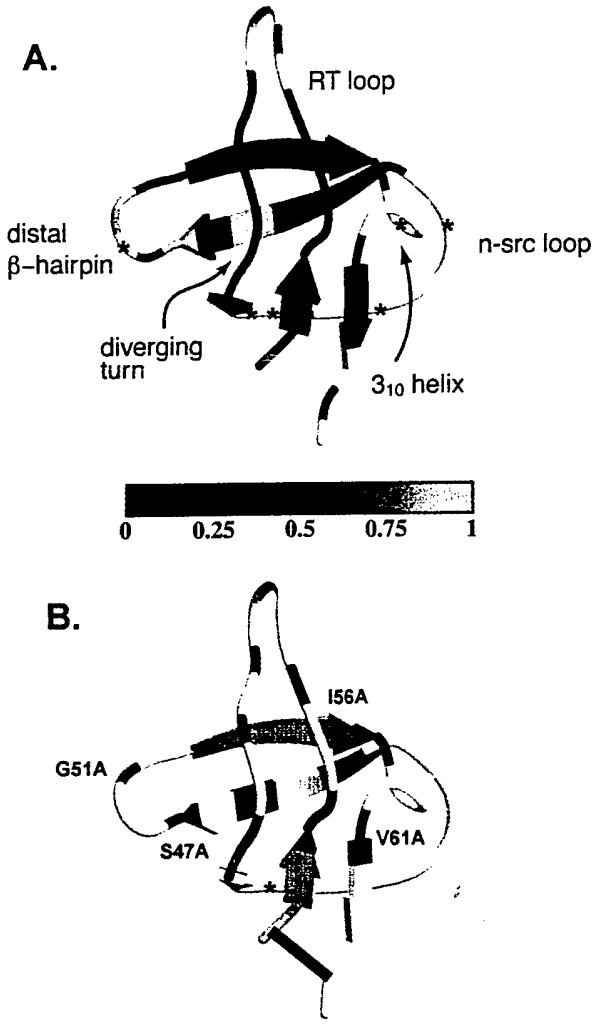


Figure 1. Structure in the transition state of (a) wt, (b) NC terminal crosslink. Color scheme is continuous from yellow ($\Phi_F = 1$) to red ($\Phi_F = 0.5$) to blue ($\Phi_F = 0$). Asterisks indicate negative Φ_F values, which suggest the involvement of these residues in non-native interactions in the transition state. Graphics were generated with Molscript[28] and Raster3d.[29,30]

parts of the molecule to overcome the loss of entropy during folding. Here, we examine the effects on the transition state of disulfide crosslinking the distal β-hairpin and circularizing (linking the N and C termini) the protein. These modifications were previously characterized kinetically in their reduced and oxidized forms to test backbone conformational ordering in the transition state.[12] Crosslinking the distal hairpin increased the folding rate 30-fold without affecting the unfolding rate, suggesting that this structural element is as conformationally constrained in the transition state as in the native state. Crosslinking the N and the C termini stabilized the protein significantly both by increasing the folding rate and by decreasing the unfolding rate, indicating that the two termini are not fully interacting in the transition state. Here, we perform mutational analysis to determine if these modifications affect the distribution of structure in the transition state. By deleting parts of individual residues (as in mutations to an alanine residue) and then assessing the effect of the mutation on stability and folding kinetics, we can gain site-specific information about structure at the rate-limiting step.[13] The degree of structure formation around each residue in the transition state can be conveniently represented by ϕ_F values, defined as $\Delta\Delta G_{U-\ddagger}/\Delta\Delta G_{U-F}$ (see Materials and Methods). In the case of the distal hairpin crosslink, we investigate whether it allows overcoming of the entropic cost of ordering earlier in the folding reaction and thus makes other parts of the transition state less structured. In the case of the terminal crosslink, we examine if significantly decreasing the entropic barrier to folding and making the protein topologically symmetric causes its transition state to become delocalized with all residues contributing equally, or whether it remains structurally polarized. A similar circularization experiment was performed on chymotrypsin inhibitor 2 (CI2),[14] however, circularization did not affect its transition state probably because it is largely delocalized even in the wild-type protein.[15] Our results suggest that, at least for the src SH3 domain, the transition state ensemble can be shifted when the native topology is significantly perturbed as in circularization, but not by stabilization of the existing nucleus.

Results

Disulfide crosslinking of the distal β-hairpin

Covalent crosslinking of the distal β-hairpin is expected to decrease the entropy of the denatured state and stabilize intrahairpin interactions. If both chain entropy and energy are smoothly varying functions of the degree of ordering and the position of the transition state is determined by their imperfect cancellation, then such a change should alter the position of the transition state. Consistent with the Hammond postulate,[16] destabilizing the denatured state would shift the position of the transition state closer to the denatured state and result in lower Φ_F values in regions other than the distal hairpin. In contrast, if the energy decreases abruptly when a large number of contacts form simultaneously, the transition state would be less sensitive to changes in interaction strengths and it would be effectively "locked". In that case, we would expect that crosslinking the distal hairpin will increase the folding rate, but structure in the transition state will remain the same. The distal hairpin was previously crosslinked by mutating residues W43 and S58 to cysteine residues and forming a disulfide bridge between them under oxidizing conditions.[12] Here, we perform Φ value analysis on several mutants throughout the crosslinked protein (denoted SS) to determine whether

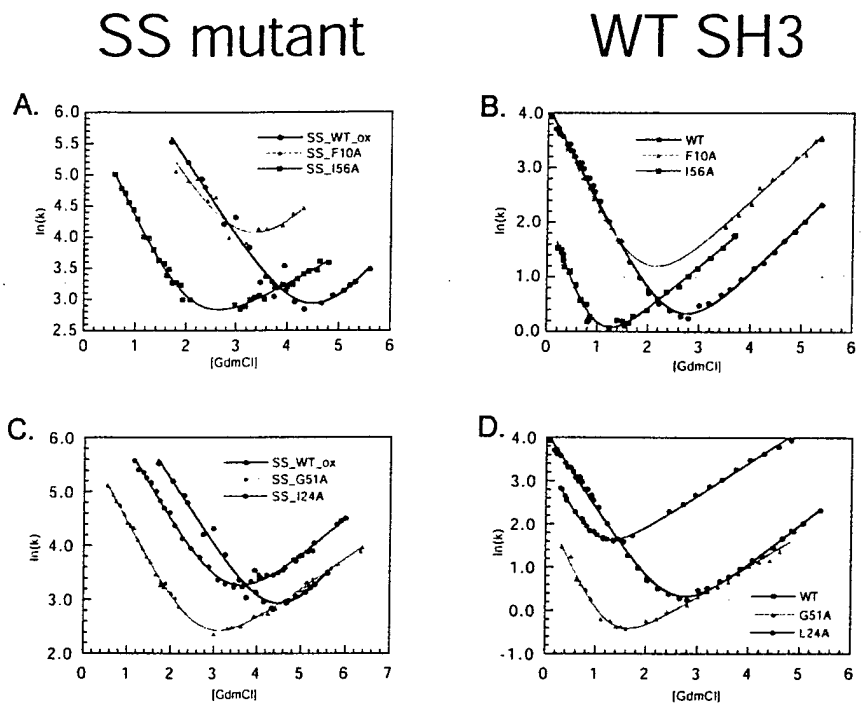


Figure 2. Kinetic analysis of mutants in the context of the distal hairpin SS crosslink. Rates of folding and unfolding were measured using stopped flow fluorescence at 295 K. Continuous lines represent the best fit to the experimental data (Kaleidagraph).

structure in the transition state has changed (F10A and A12G (N-terminal strand), D15A and L24A (RT loop), G29A and E30A (diverging turn), I34A (n-src loop) and G51A and I56A (distal β-hairpin)). The relative effect of the mutations on the rate of folding and unfolding is very similar in the context of the crosslinked protein and in the wild-type (Figure 2 and Table 1) resulting in similar Φ_F values. Even though the distal hairpin is covalently crosslinked at its base, G51A retains a Φ_F value close to 1, suggesting that this residue plays an important role in organizing structure at the turn (the glycine may allow the sidechain oxygens of S47 and T50 to hydrogen bond with neighboring backbone amides, positioning them to interact with the diverging turn). The only mutations that exhibit different behavior in the wt and crosslinked protein are L24A and I34A, but these effects can be attributed to changes in local structure around the disulfide crosslink. In the wild-type (wt) protein, I34A destabilizes the transition state more than the native state, suggesting non-native structure in the transition state, however, in the crosslinked mutant the I34A mutation has an intermediate ϕ_F value, perhaps because the replacement of the large W43 sidechain by cysteine destabilizes non-native structure in this region. L24A, on the other hand, shows an increase in Φ_F value in the crosslinked mutant from 0.26 to 0.42 indicating that the diverging turn is slightly better packed onto the distal hairpin in the transition state. Overall, however, the transition state of the crosslinked protein involves the same

Table 1. Kinetic parameters for the SS mutants

Mutant	$\ln(k_f)^{1.5\ M}$	$\ln(k_u)^{5\ M}$	m_f	m_u	$\Delta\Delta G_U$	Φ_F^{SS}	$\Phi_F^{WT\ b}$
SS_WT[a]	5.83	2.83	0.716	0.600	-	-	-
SS_F10A	5.50	4.98	0.723	0.519	−1.45	0.13	0.10
SS_D15A	5.70	3.90	0.666	0.358	−0.721	0.13	−0.22
SS_L24A	5.12	3.80	0.803	0.404	−0.985	0.42	0.26
SS_G29A	4.42	5.82	0.630	0.638	−2.58	0.32	0.44
SS_E30A	4.10	4.35	0.581	0.482	−1.91	0.53	0.62
SS_G51A	3.64	3.19	0.911	0.332	−1.50	0.86	1.06
SS_I56A	3.36	3.71	1.06	0.276	−1.96	0.74	0.71

k_f is reported in 1 M guanidine, while k_u is in 6 M guanidine to avoid extrapolation; m_f and m_u are the dependences of the folding and the unfolding rates, respectively, on Gnd. Typical errors for the kinetic measurements are 2-20% as reported by Riddle *et al.*[8]

[a] Kinetic data for this mutant was published previously by Grantcharova *et al.*[12]

[b] Φ_F values taken from the paper by Riddle *et al.*[8]

structural elements as that of the wild-type SH3 domain. We can conclude that even though formation of the distal β-hairpin is required for the overcoming of the activation barrier, it is not sufficient, even when it is largely stabilized. The rate-limiting step involves bringing the distal β-hairpin and the diverging turn together to form a three-stranded β-sheet. Thus, stabilization of this element speeds folding, but does not alter the transition state ensemble.

Disulfide crosslinking of the N and C termini

Theoretical models of the transition state for folding emphasize the balance between loss of configurational entropy and formation of stabilizing interactions in determining which part of the molecule folds first.[17–19] The combination of structural elements in the protein that can bury the most surface area while losing the least amount of configurational entropy may nucleate folding. In the modeled free energy landscape for the src SH3 domain[2] there is only one set of segments (the distal hairpin and the diverging turn) which can associate with sufficient number of favorable contacts to compensate for the loss in entropy; all other pairings are entropically too costly and poorly populated to lead to productive folding. In particular, the two terminal strands, which form a sheet in the native state, were found to be completely unstructured in the transition state due to their large sequence separation. Our strategy here is to connect the termini and examine how the distribution of structure in the transition state changes. A circularized version of the src SH3 domain (denoted NC protein) was previously constructed by mutating both residues T9 and S64 to cysteine residues and forming a disulfide bridge between them under oxidizing conditions.[12] Crosslinking makes the topology of the protein symmetric and entropically there is no reason why one three-stranded sheet would form first over the other. One prediction is that circularization would greatly reduce structural polarization because it will offer alternative routes for folding. Another possibility is that the same folding nucleus will be maintained because the interactions present in it are inherently more favorable. Such a breakdown in symmetry is seen in protein L which is topologically symmetric and yet one part of the molecule is preferentially structured at the transition state.[20,21] Distinguishing between these two possibilities addresses the relative importance of variations in interaction energies and chain entropy in determining the folding transition state.

In order to determine the effect of circularization on the transition state we examined the effect of mutants in the context of the crosslinked protein. A total of 14 mutants were designed to probe different regions of the transition state: A12G, L13A, Y16A and D23A (first strand) and RT loop; F26A and G29A (diverging turn); I34A and W43A (n-src loop); A45G, S47A and G51A (distal hairpin); I56A and P57A (3_{10} helix); V61A (C-terminal strand). Kinetic analysis (Figure 3 and Table 2) reveals that there are clear differences between some of the mutants in the NC protein and the corresponding mutations in the wt.[8] The most significant changes are observed in the Φ_F values of residues in the distal hairpin. A45G, S47A and G51A affect both the rates of folding and unfolding in the NC mutant, while in the context of the wt protein their Φ_F values were all 1 (Figure 3(a) and (b)). Since we are not certain of the homogeneity of

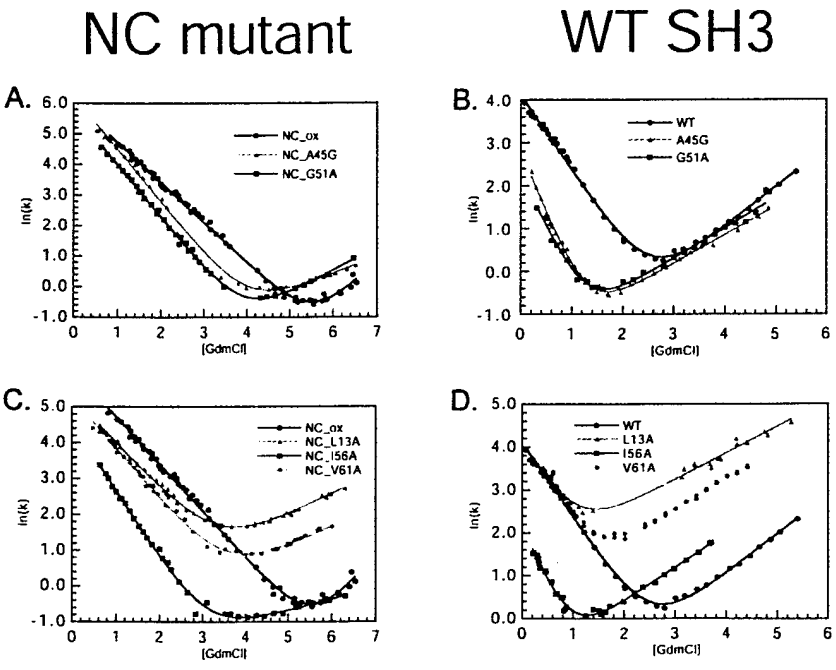


Figure 3. Kinetic analysis of mutants in the context of the NC terminal crosslink. Rates of folding and unfolding were measured using stopped flow fluorescence at 295 K. Continuous lines represent the best fit to the experimental data (Kaleidagraph).

Table 2. Kinetic parameters for the NC crosslink mutants

Mutant	$\ln(k_f)^{1\ M}$	$\ln(k_u)^{6\ M}$	m_f	m_u	$\Delta\Delta G_U$	Φ_F^{NC}	$\Phi_F^{WT\ b}$
NC_WT[a]	4.73	−0.597	0.766	0.800	-	-	-
NC_A12G	4.41	2.76	0.874	0.637	−2.16	0.09	0.05
NC_L13A	4.00	2.58	0.740	0.340	−2.29	0.19	−0.03
NC_Y16A	4.01	3.28	0.752	0.431	−2.69	0.16	0.03
NC_D23A	4.64	0.305	0.760	0.500	−0.580	0.09	0.13
NC_F26A	3.26	1.13	0.812	0.370	−1.87	0.46	0.40
NC_G29A	3.46	1.78	0.780	0.500	−2.14	0.35	0.44
NC_I34A	3.66	−1.22	1.082	0.216	−0.263	[c]	[c]
NC_W43A	3.97	1.75	0.830	0.337	−1.82	0.25	0.15
NC_A45G	4.50	0.436	1.00	0.322	−0.737	0.18	1.20
NC_S47A	3.95	0.400	0.990	0.309	−1.04	0.44	0.95
NC_G51A	3.99	0.551	0.994	0.437	−1.11	0.39	1.06
NC_I56A	2.67	−0.353	1.10	0.190	−1.35	0.89	0.71
NC_P57A	3.78	1.95	0.770	0.550	−2.05	0.27	0.24
NC_V61A	3.84	1.66	0.821	0.315	−1.84	0.28	−0.06

k_f is reported in 1 M guanidine, while k_u is in 6 M guanidine to avoid extrapolation; m_f and m_u are the dependences of the folding and the unfolding rates, respectively, on Gnd. Typical errors for the kinetic measurements are 2-20% as reported by Riddle *et al.*[8]

[a] Kinetic data for this mutant was published by Grantcharova *et al.*[12]

[b] Φ_F values taken from the paper by Riddle *et al.*[8]

[c] Mutation decreases both k_f and k_u.

the transition state the intermediate Φ_F values can either mean that the interactions in which these residues participate are not completely formed in the transition state, or that the transition state ensemble consists of some conformations in which the distal hairpin is formed and some in which the hairpin is disordered. On the other hand, mutations which probe formation of the hairpin newly created by the crosslink and the region around the 3_{10} helix have increased Φ_F values, suggesting that these residues now contribute to stabilization of the transition state. I56, an integral part of the hydrophobic core exhibits an increase in Φ_F value from 0.7 to 0.89 (Figure 3(c) and (d)). In the n-src loop, mutation of the buried W43 to alanine had no effect on the rate of folding in the wt context, but in the NC protein it has a Φ_F value of 0.25. In a similar way, V61 (C-terminal strand), which takes part in the hydrophobic core, and L13 (N-terminal strand), which interacts with the C-terminal strand on the solvent exposed side, both have increased Φ_F values upon mutation from 0 to 0.28 and 0.19, respectively. Other mutations in the first strand and the RT loop (A12G and D23A) exhibit roughly the same Φ_F values in the NC mutant as in the WT protein (Φ_F values close to 0), suggesting that despite the crosslink this region remains unstructured in the transition state. F26A and G29A in the diverging turn also preserve their intermediate Φ_F values in the NC mutant.

Taken together, these data suggest that the transition state of the circularized protein is significantly different from that of the wt protein (Figure 1(a) and (b); Figure 4(b)). The transition state of the WT SH3 domain is highly polarized with the distal hairpin and the diverging turn almost fully ordered in a three-stranded sheet, and the termini disordered. In contrast, the circularized protein appears to have a more delocalized transition state with a prevalence of intermediate Φ_F values in most of the structural elements probed. The distal hairpin, however, is still more ordered than other hairpins in the protein, probably because it has the highest density of intrahairpin interactions. It is interesting that one residue, I56 (central hydrophobic core residue), stands out with a Φ_F value close to 1 and therefore can be viewed as the nucleus around which structure consolidates. We can surmise that because of the circular topology of the protein, hairpin formation is not as important in the NC protein as it is in the wt. Instead, it appears that hydrophobic collapse, rather than local β-hairpin and sheet formation,

Table 3. Kinetic parameters for WT and mutants in 0.4 M sodium sulfate

Mutant	$\ln(k_f)^{0.5\ M}$	$\ln(k_u)^{5\ M}$	m_f	m_u	$\Delta\Delta G_U$	Φ_F^{sulf}	$\Phi_F^{WT\ b}$
WT_sulf	4.31	0.867	0.695	0.805	-	-	-
F10I_sulf	4.22	3.65	0.702	0.580	−1.68	0.03	−0.05
L44A_sulf	2.87	2.66	1.24	0.474	−1.89	0.45	0.54
G51A_sulf	2.33	1.12	0.959	0.519	−1.31	0.89	1.06
I56A_sulf	2.06	1.70	1.14	0.468	−1.81	0.73	0.71

k_f is reported in 0.5 M guanidine, while k_u is in 5 M guanidine to avoid extrapolation; m_f and m_u are the dependences of the folding and the unfolding rates, respectively, on Gnd. Typical errors for the kinetic measurements are 2-20% as reported by Riddle *et al.*[8]

[a] Φ_F values taken from the work by Riddle *et al.*[8]

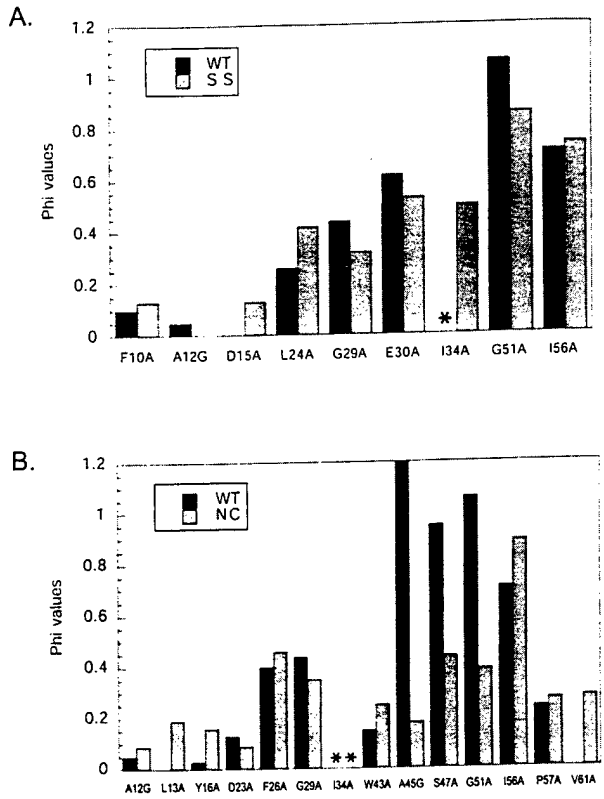


Figure 4. Comparison of Φ_F values for mutated residues in the wt and the (a) SS crosslink and (b) NC crosslink. Asterisk indicates a negative Φ_F value, which suggests the involvement of this residue in non-native interactions in the transition state.

drives the early stages of folding of the circularized protein, due to the decreased cost of bringing together residues distant in the chain.

Global stabilization and structure in the transition state

Probing the effect of a globally stabilizing agent on the rate-limiting step in folding provides another way to examine the robustness of the transition state. Sodium sulfate stabilizes proteins, presumably by its preferential hydration of water, therefore facilitating hydrophobic collapse.[22] Its effect on the kinetics of the src SH3 domain is to increase the folding rate and to decrease the unfolding rate (Table 3), indicating that protein desolvation occurs both before and after the transition state. It also decreases the denaturant dependence of the folding rate (i.e. m_f value), suggesting that it makes the denatured state more compact. We performed kinetic analysis in the presence of 0.4 M sodium sulfate of several mutants, in the wt context, which in the absence of sodium sulfate cover the full range of Φ_F values: F10I (Φ_F value of 0); G51A (Φ_F value of 1); L44A and I56A (intermediate Φ_F values). All mutants conserve their Φ values in the presence of sodium sulfate (Table 3), suggesting that the transition state ensemble has not been changed by addition of the salt. This is a further confirmation that the transition state for folding is determined by deep features of the SH3 energy landscape. Similar experiments with the α-spectrin SH3 domain show that variations in pH do not affect transition state structure.[7]

Discussion

Changing the structure of transition state ensembles

Point mutagenesis, which probes residue-specific interactions, and covalent modifications (glycine loop insertions and disulfide crosslinking), which test long-range order in the transition state revealed the conformationally restricted nature of the transition state ensemble of the src SH3 domain.[8,12] Here, we have explored in more detail the free-energy landscape of this protein and have determined how changes in chain configurational entropy and interaction strengths affect the distribution of structure in the transition state.

Covalent crosslinking is a convenient way of altering the entropic cost of contact formation. It reduces the average sequence separation between interacting residues (i.e. contact order (CO)) and is therefore expected to increase the rate of folding, as observed in the very good correlation between CO and rate of folding for all the characterized two state folding proteins.[23] Both the NC and SS-crosslinked mutants fall within the spread observed for natural proteins on the CO *versus* log(k) plot (data not shown). It should be noted, however, that the distal hairpin crosslink causes a larger increase in folding rate than the NC-terminal crosslink even though the contact order of the NC protein is smaller. This is because the folding rate is sensitive to chain entropy loss in the transition state rather than the native state. The distal loop crosslink reduces the entropy of the denatured state dramatically but has essentially no effect on that of the transition state (the loop is already formed), while the NC crosslink reduces the entropy of both the denatured state and the transition state. The overall correlation between contact order and folding rate suggests that on average the distribution of contacts (contact order) in the folding transition state follows that of the native structure, but in any specific case, the effect of a crosslink on the folding rate will depend on the transition state structure and not solely on the reduction in the native state contact order.

The finding that stabilization of local structure by distal hairpin crosslinking (Table 1, Figure 4(a)) and global stabilization by sodium sulfate (Table 3) do not alter the placement of the transition state along the reaction coordinate (as judged by Φ_F value distributions) indicates that there are some deep features in the energy landscape which are not altered by such changes. These results are con-

sistent with experiments on other SH3 domains. The distantly related src and α-spectrin SH3 domains exhibit very similar transition states,[7,8] and stabilizing mutations[11] and changes in pH[7] do not seem to affect transition state structure of the spectrin SH3 domain. It appears then, that SH3 domains allow quite large variations in sequence and experimental conditions with no change to the transition state probably because there are no alternative structural elements that can be sufficiently stabilized to become folding nuclei. On the other hand, modifying the topology of the protein, as in circularization, can significantly change the free energy landscape to favor alternative routes for folding. Similar conclusions were drawn from the circular permutation experiments on the α-spectrin SH3 domain.[10] Connecting the wt termini with a small peptide linker and introducing a cut in the distal hairpin resulted in a shift in the structure of the transition state towards the n-src loop and the hairpin formed by the old termini; the former distal hairpin was completely disordered at the rate-limiting step. (In contrast, circular permutations that did not involve the distal loop β-hairpin did not appear to change the folding transition state.) Therefore, shifts in transition state structure can occur when formerly distant elements are covalently linked to reduce the entropic cost of their interaction.

It should be noted that SH3 domains have more polarized folding transition states than other small proteins (CI2, ACBP, AcP, FKBP12). Therefore, changes in the structure of the folding transition state are more evident for the SH3 domains than for proteins with more delocalized folding transition state ensembles. A particularly well studied example of a protein with a more delocalized folding transition state is chymotrypsin inhibitor 2 (CI2), only one residue has a Φ_F value greater than 0.5. Drastic changes in the topology of CI2 through circular permutation or circularization[14] have relatively little effect on the folding transition state.

While topology plays an important role in determining a protein's folding mechanism, the distribution of interaction energies throughout the protein also affects structure in the transition state. Recent experiments demonstrate that the transition state conservation observed for sequence homologs of the SH3 domain does not hold for structural homologs. For example, drastic mutagenesis, which weakens the interaction energies throughout the protein can make the transition state delocalized. A sequence simplified mutant of the src SH3 domain made predominantly of five amino acid residues (I, K, E, A, G) was found to have a more delocalized transition state (distal hairpin is not fully formed), most likely because the interactions stabilizing the wt SH3 transition state are not strong enough in the simplified mutant to overcome the loss in entropy and residues from other parts of the protein have to participate (Q. Yi and D.B., unpublished results). Furthermore, the presence of destabilizing features in a particular structural element might be required for functional reasons. This results in a preferential switch in the structured parts of the transition state to other regions with more favorable interactions. PsaE, a structural homolog of the SH3 domain, has a large loop insertion at the distal hairpin (13 residues) required for its function in the photosynthetic center I of cyanobacteria.[24] The larger entropic cost of forming stabilizing interactions makes the transition state delocalized with high Φ_F values distributed throughout the protein (P. Bowers and D.B., unpublished results). Sso7d, a DNA binding protein from *Sulfolobus solfataricus*,[1] is another structural homolog of the SH3 domain. Its distal hairpin contains three glycine residues at the turn and two more in the β-strands required for function, and is not well ordered early in folding (R. Guerois and L. Serrano, personal communication). The n-src hairpin, on the other hand, is the most regular element of structure with a favorable hydrophobic/hydrophilic pattern along the strands and a canonical type I turn. The burial of hydrophobic surface area between the n-src loop and the C-terminal helix further favors these elements as a folding nucleus. Therefore, the topology of the SH3 fold appears to allow several alternative routes of folding.

Another example of a simple system in which the effects of topology and local structural propensity on the transition state have been examined is the GCN4-p1 coiled coil. Mutational analysis indicated that the folding of the dimeric coiled coil occurs *via* multiple pathways.[25] Variations in helical propensity along the helix can favor one pathway over the others (e.g. the C terminus of the GCN4 coiled coil has a higher helical propensity and has been shown to form early in folding).[26] Destabilizing one part of the helix (with A to G mutations) channels folding to the alternative pathways. However, crosslinking the two helices to form a monomer abolishes the symmetry, making it entropically more favorable for folding to start at the part of the helix proximal to the tether, even if it has the lowest helical propensity.[25] In the monomeric version of the coiled coil, the topological constraints on the chain effectively limit the number of folding pathways to one and make the transition state less sensitive to variations in secondary structure.

A similar dependence of the folding mechanism on the stability of individual structural elements is observed in two proteins with symmetrical topology: protein L and protein G (an α-helix packed against two β-hairpins forming a sheet). In the transition state of protein L the first hairpin packs against the α-helix, while in protein G the second hairpin is more structured.[20,21] The choice of hairpin appears to depend on the intrinsic stability of the hairpins. In protein L, the first hairpin has more favorable side-chain:main-chain hydrogen bonds, while the second hairpin contains three consecutive residues with positive φ angles. In protein G, on the other hand, the second hairpin has an

extensive hydrogen bond network. Using computational protein design methods, the order of events in the folding of protein L and protein G can be switched by selectively stabilizing the hairpin normally formed late in folding (S. Nauli, B. Kuhlman & D.B., unpublished results).

The ability to change the transition state for folding tests our understanding of the factors contributing to its formation and specificity. Our results with the circularized src SH3 domain and the experimental studies on other proteins highlight the interplay of topologic constraints and contact energy heterogeneity in determining the structure of the transition state ensemble.

Materials and Methods

Mutagenesis

Point mutagenesis was accomplished using the Quick Change Site-Directed mutagenesis kit (Stratagene, La Jolla, CA). Plasmids harboring the mutations were transformed into BL21 cells, and protein was overexpressed and purified.[9] The His·Tag® was not removed for the purpose of this study. All mutants were sequenced to ensure that the mutagenesis was successful and the purified proteins were analyzed by mass spectrometry to confirm that each mutation was the expected one.

Disulfide crosslinking

The design of the SS and NC crosslink mutants was described by Grantcharova *et al.*[12] For all the mutants disulfide bonds were oxidized in the presence of 20 mM potassium ferricyanide $K_3Fe(CN)_6$ for ten minutes at room temperature. Reactions were performed in the dark because $K_3Fe(CN)_6$ is light sensitive. Disulfide formation was confirmed using Ellman's reagent.

Biophysical analysis

Protein solutions (100 μM) were made in 50 mM sodium phosphate (pH 6). For the experiments in sodium sulfate, 0.4 M sulfate was added to the solutions. The kinetics of folding and unfolding were followed by tryptophan fluorescence on a Bio-Logic SFM-4 stopped-flow instrument at 295 K. The unfolding reaction for the wt protein was previously determined to behave as a two-state process,[27] and the kinetic and equilibrium data for the mutants were fit to a two-state model. For each mutant the free energy of folding is calculated as:

$$\Delta G_{U-F} = RT \ln(k_f/k_u)$$

where k_f and k_u are the rates of folding and unfolding, respectively, in the absence of denaturant. The difference in the free energy of folding ($\Delta\Delta G_{U-F}$) and in the folding activation energy ($\Delta\Delta G_{U-\ddagger}$) between the wt protein and each mutant are calculated as:

$$\Delta\Delta G_{U-F} = RT(\ln(k_f^{wt}/k_f^{mut}) + \ln(k_u^{mut}/k_u^{wt}))$$

and

$$\Delta\Delta G_{U-\ddagger} = RT \ln(k_f^{wt}/k_f^{mut})$$

where k_f and k_u are the rates of folding and unfolding, respectively, at denaturant concentrations experimentally accessible for that mutant. This method avoids the extrapolation of k_f and k_u to 0 M denaturant and therefore does not rely on the accurate determination of the m_f and m_u values (the denaturant dependence of k_f and k_u, respectively).

The parameter Φ_F is defined as:

$$\Phi_F = \Delta\Delta G_{U-\ddagger}/\Delta\Delta G_{U-F}$$

and is interpreted as the fraction of the mutated residue's interactions that are formed in the transition state. A Φ_F value of 1 indicates that all of a residue's interactions are formed in the transition state, whereas a Φ_F of 0 means that the residue does not make stabilizing interactions in the transition state.[13]

Acknowledgments

We thank Raphael Guerois and Luis Serrano for communicating to us their unpublished results on the Sso7d protein. We are grateful to members of the Baker group for their useful comments on the manuscript. This work was supported by a grant from the NIH and Young Investigator awards to D. B. from the NSF and the Packard Foundation.

References

1. Baumann, H., Knapp, S., Lundback, T., Ladenstein, R. & Hard, T. (1994). Solution structure and DNA-binding properties of a thermostable protein from the archaeon Sulfolobus solfataricus. *Nature Struct. Biol.* **1**, 808-819.
2. Alm, E. & Baker, D. (1999). Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl Acad. Sci. USA*, **96**, 11305-11310.
3. Debe, D. A. & Goddard, W. A., III (1999). First principles prediction of protein folding rates [In Process Citation]. *J. Mol. Biol.* **294**, 619-625.
4. Galzitskaya, O. V. & Finkelstein, A. V. (1999). A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl Acad. Sci. USA*, **96**, 11299-11304.
5. Muñoz, V. & Eaton, W. A. (1999). A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl Acad. Sci. USA*, **96**, 11311-11316.
6. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Struct. Biol.* **6**, 1005-1009.
7. Martinez, J. C. & Serrano, L. (1999). The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nature Struct. Biol.* **6**, 1010-1016.
8. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. & Baker, D. (1999). Experiment and theory highlight role of native state topology in SH3 folding. *Nature Struct. Biol.* **6**, 1016-1024.
9. Grantcharova, V. P., Riddle, D. S., Santiago, J. V. & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain [see comments]. *Nature Struct. Biol.* **5**, 714-720.

10. Viguera, A. R., Serrano, L. & Wilmanns, M. (1996). Different folding transition states may result in the same native structure. *Nature Struct. Biol.* **3**, 874-880.
11. Martinez, J. C., Pisabarro, M. T. & Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state [see comments]. *Nature Struct. Biol.* **5**, 721-729.
12. Grantcharova, V. P., Riddle, D. S. & Baker, D. (2000). Long-range order in the src SH3 folding transition state [In Process Citation]. *Proc. Natl Acad. Sci. USA*, **97**, 7084-7089.
13. Fersht, A. R. (1995). Characterizing transition states in protein folding: an essential step in the puzzle. *Curr. Opin. Struct. Biol.* **5**, 79-84.
14. Otzen, D. E. & Fersht, A. R. (1998). Folding of circular and permuted chymotrypsin inhibitor 2: retention of the folding nucleus. *Biochemistry*, **37**, 8139-8146.
15. Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260-288.
16. Matouschek, A. & Fersht, A. R. (1993). Application of physical organic chemistry to engineered mutants of proteins: Hammond postulate behavior in the transition state of protein folding. *Proc. Natl Acad. Sci. USA*, **90**, 7814-7818.
17. Clementi, C., Jennings, P. A. & Onuchic, J. N. (2000). How native-state topology affects the folding of dihydrofolate reductase and interleukin-1beta. *Proc. Natl Acad. Sci. USA*, **97**, 5871-71876.
18. Clementi, C., Nymeyer, H. & Onuchic, J. N. (2000). Topological and energetic factors: what determines the structural details of the transition state ensemble and "Enroute" intermediates for protein Folding? An investigation for small globular proteins. *J. Mol. Biol.* **298**, 937-953.
19. Onuchic, J. N., Nymeyer, H., Garcia, A. E., Chahine, J. & Socci, N. D. (2000). The energy landscape theory of protein folding: insights into folding mechanisms and scenarios. *Advan. Protein Chem.* **53**, 87-152.
20. McCallister, E. L., Alm, E. & Baker, D. (2000). Critical role of beta-hairpin formation in protein G folding. *Nature Struct. Biol.* **7**, 669-673.
21. Kim, D. E., Fisher, C. & Baker, D. (2000). A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* **298**, 971-984.
22. Timasheff, S. N. & Arakawa, T. (1989). Stabilization of protein structure by solvents. In *Protein Structure: A Practical Approach* (Creighton, T. E., ed.), pp. 331-335, IRL Press, Oxford.
23. Plaxco, K. W., Simons, K. T. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985-994.
24. Falzone, C. J., Kao, Y. H., Zhao, J., Bryant, D. A. & Lecomte, J. T. (1994). Three-dimensional solution structure of PsaE from the cyanobacterium Synechococcus sp. strain PCC 7002, a photosystem I protein that shows structural homology with SH3 domains. *Biochemistry*, **33**, 6052-6062.
25. Moran, L. B., Schneider, J. P., Kentsis, A., Reddy, G. A. & Sosnick, T. R. (1999). Transition state heterogeneity in GCN4 coiled coil folding studied by using mutlisite mutations and crosslinking. *Proc. Natl Acad. Sci. USA*, **96**, 10699-10704.
26. Zitzewitz, J. A., Ibarra-Molero, B., Fishel, D. R., Terry, K. L. & Matthews, C. R. (2000). Preformed secondary structure drives the association reaction of GCN4- p1, a model coiled-coil system. *J. Mol. Biol.* **296**, 1105-1116.
27. Grantcharova, V. P. & Baker, D. (1997). Folding dynamics of the src SH3 domain. *Biochemistry*, **36**, 15685-15692.
28. Kraulis, P. J. (1991). MOLSCRIPT: A Program to Produce Both Detailed and Schematic Plots of Protein Structures. *J. Appl. Crystallog.* **24**, 946-950.
29. Bacon, D. J. & Anderson, W. F. (1988). A fast algorithm for rendering space-filling molecule pictures. *J. Mol. Graph.* **6**, 219-220.
30. Merritt, E. A. & Bacon, D. J. (1997). Raster3D: Photorealistic Molecular Graphics. *Methods Enzymol.* **277**, 505-524.

Edited by C. R. Matthews

(Received 27 July 2000; received in revised form 16 November 2000; accepted 21 November 2000)

Mechanisms of protein folding

Viara Grantcharova*, Eric J Alm†, David Baker† and Arthur L Horwich‡

The strong correlation between protein folding rates and the contact order suggests that folding rates are largely determined by the topology of the native structure. However, for a given topology, there may be several possible low free energy paths to the native state and the path that is chosen (the lowest free energy path) may depend on differences in interaction energies and local free energies of ordering in different parts of the structure. For larger proteins whose folding is assisted by chaperones, such as the *Escherichia coli* chaperonin GroEL, advances have been made in understanding both the aspects of an unfolded protein that GroEL recognizes and the mode of binding to the chaperonin. The possibility that GroEL can remove non-native proteins from kinetic traps by unfolding them either during polypeptide binding to the chaperonin or during the subsequent ATP-dependent formation of folding-active complexes with the co-chaperonin GroES has also been explored.

Addresses
*Center for Genomics Research, Harvard University, Cambridge, MA 02138, USA
†Department of Biochemistry and Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA
‡Department of Genetics and Howard Hughes Medical Institute, Yale School of Medicine, New Haven, CT 06510, USA

Current Opinion in Structural Biology 2001, 11:70–82

0959-440X/01/$ – see front matter

Abbreviations

AcP	acylphosphatase
Ada2h	activation domain of procarboxypeptidase
CO	contact order
EDTA	ethylenediamine tetra-acetic acid
GFP	green fluorescent protein
MDH	malate dehydrogenase
Rubisco	ribulose-1,5-bisphosphate carboxylase-oxygenase
SH	Src homology
TFE	trifluoroethanol

Introduction

Two aspects of protein folding mechanisms are considered in this review: recent insights into the folding behavior of small two-state folding proteins and the action of the chaperonin GroEL in assisting the folding of larger proteins.

Folding of small proteins

The past several years have witnessed a rapid increase in the amount of experimental data on the folding of small single-domain proteins. Comparison of results on sets of both homologous and unrelated proteins has provided considerable insight into the determinants of the folding process. In this part of the review, we present simple models that incorporate recent experimental findings and appear to capture the broad outlines of the folding process. An important feature of these models is that the folding free energy landscape is dominated by the trade-off between the unfavorable loss in configurational entropy upon folding and the gain in attractive native interactions; non-native interactions are assumed not to play a significant role. As will be discussed first, recent results suggest a picture in which several different routes through the free energy landscape with roughly equivalent free energy barriers can be consistent with the overall topology (low-resolution structure) of a protein and sequence changes can, by lowering or raising one barrier relative to another, produce significant changes in the transition-state ensemble without large changes in the folding rate. Because our recent articles have probably overly emphasized the role of native state topology [1–3], we shall subsequently focus our attention on several examples that illustrate how variations in local free energies of ordering can modulate the folding process.

We begin by considering a zeroth order model in which all native interactions in a protein are equally favorable (i.e. homogeneous contact model). In such a model, the free energy cost of forming different contacts in a protein depends solely on the entropic cost of restricting the chain to allow the contact. This entropic cost increases with increasing sequence separation between the interacting residues, as more of the chain must be constrained in order to form the contact. When many of the contacts in a protein are between residues distant in the primary sequence, a large portion of the chain must be ordered before even a few favorable contacts can form, leading to a large folding free energy barrier. Conversely, when interacting residues are close in the protein sequence, the entropic cost of chain ordering is partially compensated by the formation of contacts earlier in the folding process, leading to a smaller folding free energy barrier. Therefore, in this very simple model, one expects proteins with most of their contacts between residues close in the sequence to fold faster than proteins with contacts between residues distant in the sequence.

Several years ago, we found such a relationship between folding rate and the average sequence separation between contacting residues (the contact order — CO) [1]. A considerable number of proteins have been studied in the interim period and an updated version of the plot, encompassing all the two-state folding proteins that have been kinetically characterized (Table 1), shows an even stronger correlation between CO and rate of folding (Figure 1a). The correlation is particularly remarkable because of the very wide variation in the folds and functions of these proteins. It suggests that the low-resolution structure or topology of a protein is a major determinant of the trade-off between configurational entropy loss and formation of attractive interactions, as suggested by the simple model described in the previous paragraph. The correlation also

supports the assumption that non-native interactions play a relatively minor role in shaping the folding process as, unlike native interactions, they are not expected to be related to the native structure.

In the simple zeroth order model discussed above, increasing uniformly the strength of all interactions clearly reduces the free energy barrier to folding (the unfavorable entropy of ordering is better compensated by the formation of the more favorable interactions) and the folding rate increases. Thus, for a given protein, reducing the strength of the favorable interactions (i.e. reducing stability) is expected to reduce the folding rate. Indeed, there is a nearly linear correlation between folding rate and stability for a given protein upon changes in solution conditions, most notably upon the addition of denaturant. Also, within a protein family, more stable proteins generally fold more rapidly than less stable proteins [4,5]. However, the correlation between stability and folding rate for proteins with different folds is much weaker than that between CO and folding rate, consistent with the dominant role of native state topology in determining folding rates [2].

Interestingly, there is a better correlation between the folding rate and the relative CO (average sequence separation divided by chain length) than between the folding rate and the absolute (unnormalized) CO (compare Figure 1a,b). This is somewhat unexpected as the entropic cost of contact formation is a function of the absolute CO, rather than of the relative CO, and simple models of the sort discussed above predict relationships with the absolute CO. If the improved correlation with the relative CO is borne out by further experimental data over the next several years, it may be necessary to consider models in which there is a renormalization that removes the dependence on the absolute length of the protein. An alternative possibility is that, for the proteins in this set, the stability increases with increasing length and dividing by the length accounts for the effect of stability on the folding rate, albeit in a somewhat indirect way.

We frequently encounter two questions about the contact order/folding rate correlation. First, given that the entropic cost of closing a loop in a protein is proportional to the logarithm of the loop length, shouldn't folding rates be more closely correlated to the logarithm of the CO? As shown in Figure 1c, because of the limited range of the CO values, the relationship between folding rates and log CO is nearly indistinguishable from that between folding rates and CO. Second, as the magnitude of the entropic barrier to folding depends on the CO of the folding transition-state ensemble, why is there a correlation between folding rates and the CO of the native structure? The correlation suggests that the CO of the native structure is, in turn, correlated with that of the transition-state ensemble; this is not surprising given that a reasonable fraction of the native structure is usually formed in the transition-state ensemble and that contact lengths tend to be relatively consistent within particular protein structures (in an all-helical protein, the contact lengths are consistently shorter than in a parallel β-sheet protein, for example).

Table 1

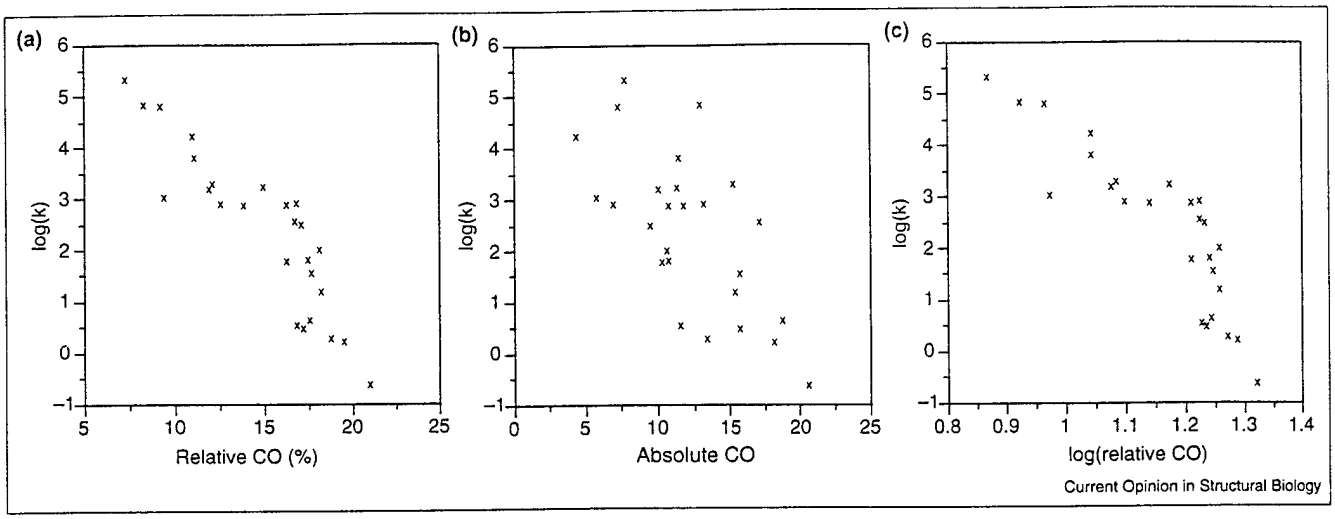
Rates of folding for two-state folding proteins.

Protein*	Log(k_f)†	CO‡ (%)	ΔG_u (kcal/mol)	Length§ (residues)	Temperature (C°)
Cyt-B_{562} [62]	5.30	7.47	10.0	106	20
Myoglobin	4.83#	8.50	8.4	154	25
λ-repressor [63]	4.78	9.37	5.6	80	20
PSBD [64]	4.20	11.20	2.2	41	41
Cyt-c [65]	3.80#	11.22	8.2	104	23
Im9 [66]	3.16	12.07	6.6	85	10
ACBP [67]	2.85	13.99	8.2	86	25#
Villin 14T [68]	3.25	12.31	9.8	126	25
N-term L9 [69]	2.87	12.74	4.5	56	25
Ubiquitin [70]	3.19	15.11	7.2	76	25
CI2 [71]	1.75	16.40	7.6	64	25
U1A [72]	2.53	16.91	9.9	102	25
Ada2h [73]	2.88	16.96	4.1	79	25
Protein G [74]	2.46	17.30	4.6	56	25
Protein L [75]	1.78	17.62	4.6	62	22
FKBP [76]	0.60	17.70	5.5	107	25
HPr [77]	1.17	18.35	4.7	85	20
MerP [78]	0.26#	18.90	3.4	72	25
mAcP [79]	−0.64	21.20	4.5	98	28
CspB [4]	2.84	16.40	2.7	67	25
TNfn3 [80]	0.46	17.35	5.3	92	20
TI I27 [80]	1.51	17.82	7.5	89	25
Fyn SH3 [5]	1.97	18.28	6.0	59	20
Twitchin [80]	0.18	19.70	4.0	93	20
PsaE[a]	0.51	17.01	1.57	69	22
Sso7d[b]	3.02	9.54	5.93	63	25

*A nonhomologous set of simple, single-domain, non-disulfide-bonded proteins that have been reported to fold via two-state kinetics under at least some conditions. Reported data and representative members of homologous families selected as previously described [1].
†Extrapolated folding rates in water. May differ from true folding rate in water (e.g. cyt-c, protein G, ubiqutin and others) due to 'roll-over' at low denaturant concentrations. ‡Calculated as previously described [1]. §Length of protein in residues from first structured residue to last. May differ from number of residues in construct characterized. #As reported previously in [2]. (a) P Bowers, D Baker, unpublished data. (b) L Serrano, personal communication.

In the simple zeroth order model, protein topology is the single most important determinant of the folding process because it determines the sequence separation and spatial arrangement of the contacting residues. Indeed, simple computational models based on the homogeneous contact picture have done reasonably well at capturing many of the overall features of protein folding rates and mechanisms [6–9]. However, there are now a number of examples in which differences in local free energies of ordering have a significant influence on the folding mechanism, particularly in cases in which several different pathways are equally consistent with the structure because of symmetry (see below). These differences may arise, for example, from particularly unfavorable local conformations that either are important for functional reasons or are compensated in the final folded structure

Figure 1



Correlation between the logarithm of the folding rate and **(a)** relative CO, **(b)** absolute CO and **(c)** log(relative CO).

by particularly favorable nonlocal interactions. Incorporation of these differences leads to a model in which the order of events in folding depends both on the overall topology and on the relative free energy of ordering different parts of the chain. Given two possible routes to the native state, which involve forming contacts between residues equally distant along the chain, the lowest free energy route is that involving the formation of the lowest free energy substructures. Such a model produces considerably better predictions of the folding rate and of the dominant features of the structure of the folding transition-state ensemble than the simple zeroth order model (see Figures 2 and 3; E Alm, A Morozov, D Baker, unpublished data).

Experimentally, the distribution of structure in the folding transition state can be determined by measuring the effect of mutations throughout the protein on the folding and unfolding rate [10]. Fersht's Φ value notation is a convenient way to summarize such data; a Φ value of one indicates that the interactions made by a residue are as ordered in the transition state as in the native state, whereas a Φ value of zero indicates that the interactions are not formed in the transition state [11]. Table 2 summarizes the general properties of the folding transition states studied so far using this kind of analysis. The following focuses on several recent examples that highlight the interplay between the native state topology and variations in local free energies of ordering in determining the folding mechanism (this is not a comprehensive summary of recent advances in protein folding studies).

GCN4 and λ repressor

The GCN4-p1 coiled coil is a particularly simple system for the detailed examination of the effects of topology and local structural propensity on the distribution of structure in the transition-state ensemble. The rate-limiting step in folding involves the association of two monomers to form a dimer in which hydrophobic residues are partially buried, but the helices are not completely formed. The C-terminal region of the helix exhibits higher helix propensity and mutations in that region have larger effects on the folding rate than mutations in the N terminus [12,13]. Interestingly, the effect of mutations on the folding rate can be altered by manipulating the helix propensity throughout the helix with the help of additional mutations. For example, once the N terminus of the helix is stabilized by two alanine substitutions, a subsequent mutation at the C terminus has a relatively small effect on folding, and when the C terminus is destabilized by a glycine substitution, a subsequent mutation at the N terminus has a much larger effect on folding than in the wild-type protein [12]. Thus, whereas in the wild-type protein the rate-limiting step appears to involve primarily the association of C-terminal portions of the two helices [13], association of the N-terminal regions can nucleate folding if the N terminus is stabilized or the C terminus is destabilized. Such malleability is expected given the symmetry of the helix — it appears that the rate-limiting step involves the pairing of helical regions of the two monomers, but whether these are C-terminal or N-terminal depends on the details of the sequence and can be perturbed by mutations that alter the helix propensity. However, when the symmetry is broken by connecting the N termini of the helices with a covalent cross-link, the portions of the helices adjacent to the (N-terminal) cross-link are largely formed and the C-terminal regions are largely disrupted in the transition state, regardless of the intrinsic helical propensities [12]. Therefore, in this system, local structural biases have some influence on the transition state when multiple folding routes are equally consistent with the overall topology because of symmetry (the dimeric

Table 2

Folding transition states characterized by mutational analysis.

Protein	Fold	Number of residues	Number of mutants	Transition state (TS) characteristics
λ Repressor	α helix	80	8	Some helices are more structured in the TS than others; multiple folding pathways were postulated because of the dramatic effect of single mutations and temperature on TS structure [14,15]
ACBP	α helix	86	26	Terminal helices come together in the TS, while the rest of the protein is involved in non-native interactions; conserved hydrophobic residues are important in the TS [67]
GCN4 coiled coil	α helix			TS for coiled-coil formation is different when the two helices are cross-linked and when they form a dimer [12,13]
Monomer		72	3	
Dimer		36/36	3	
src SH3 domain	β barrel	57	57	TS is structurally polarized, with part of the protein fully formed and the rest fully disordered; TS is conserved among distant sequence homologs [3,22]
α-Spectrin SH3 domain		62	17	
PsaE	β barrel	69	18	These proteins are structural homologs of the SH3 domain, but do not exhibit the same TS (P Bowers, D Baker, unpublished data; L Serrano, personal communication; Q Yi, D Baker, unpublished data)
Sso7d		63	24	
Simplified SH3		56	5	
src SH3 circ	β barrel	57	14	Circularization (circ) makes the TS more delocalized, whereas cross-linking (cross) of the distal hairpin leaves it unchanged [25]
src SH3 cross		57	9	
Spectrin SH3 perm1	β barrel	62	7	Permutation at the distal hairpin, but not at the RT loop, causes a shift in the structure of the TS [81]
Spectrin SH3 perm2		62	8	
TNfn3	β sandwich	92	48	Structurally polarized: a ring of core residues from the central β strands forms the folding nucleus, while the terminal strands are disordered [82]
Ada2h	α/β	81	15	The topology of this fold allows several different TSs, depending on which helix is more structured [19–21,73]
AcP	(βαββαββ)	98	26	
U1A		102	13	
S6		101	?	
Protein L	α/β	62	70	The symmetric topology of the protein allows for two possible TSs, depending on which hairpin is more stable; stabilizing the opposite hairpin leads to a switch in the transition state (protein G_Nu); ([16,17]; S Nauli, B Kuhlman, D Baker, unpublished data)
Protein G	(ββαββ)	57	19	
Protein G_Nu		57	4	
CI2	α/β	64	150	Delocalized TS, with most of the interactions only partially formed [71]
CI2 circ	α/β	64	11	Circularization (circ), circular permutation (perm) and fragmentation (frag) do not change the delocalized TS [83]
CI2 perm		64	11	
CI2 frag		40/24	23	
FKBP	α/β	107	34	[76,84]
CheY	α/β	129	34	[85]
p13suc1	α/β	113	57	[86]
Arc repressor	α/β	53	44	Delocalized TS [87]

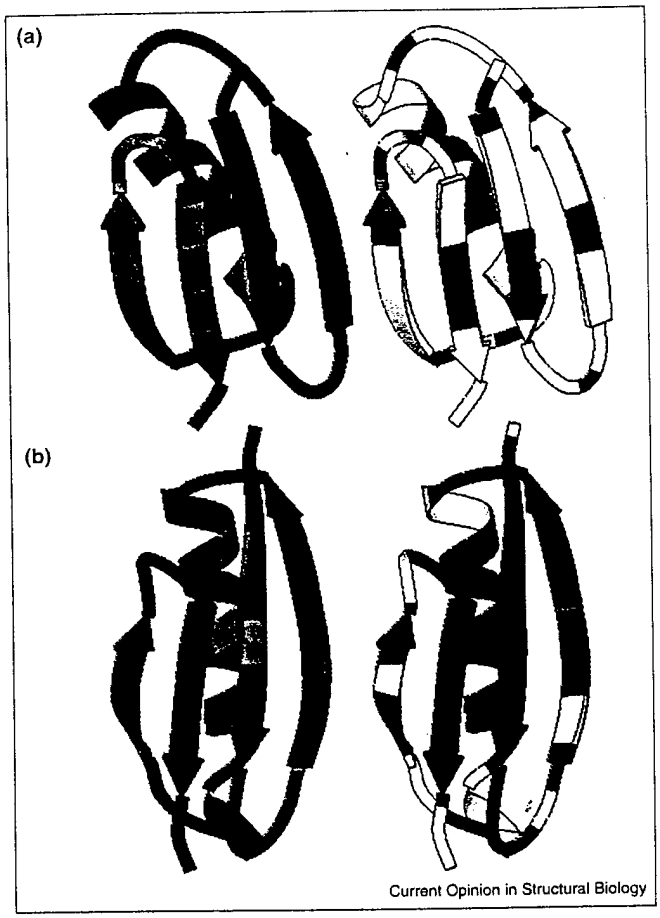
case). However, when the topology strongly favors one particular route to the native state because of the reduced entropic cost of forming more local interactions (the monomeric case), secondary structure propensities are of little consequence.

The λ repressor, another all α-helical protein, has also been postulated to fold by a number of pathways, depending on the intrinsic stability of each helix. Both point mutations [14] and temperature [15] have been shown to significantly change structure in the transition state.

Protein L and protein G

Protein L and protein G are structural homologs, but have little detectable sequence similarity. Both proteins consist of an α helix packed across a four-stranded sheet formed by two symmetrically disposed β hairpins. Remarkably, the symmetry of the fold is almost completely broken during folding: in protein L, the first hairpin is formed and the second disrupted at the rate-limiting step in folding, whereas in protein G, the second hairpin is formed and the first is disrupted [16,17] (Figure 2). Thus, despite the small size (~60 residues) of the two proteins and their topological

Figure 2



Folding transition states of **(a)** protein G and **(b)** protein L. Left, predicted phi values; right, experimental phi values. The color scheme is continuous from red ($\Phi = 1$; structured in the transition state) to blue ($\Phi = 0$; unstructured in the transition state). Sites not probed experimentally are indicated in white. Graphics were generated with MOLSCRIPT [88]. Predicted phi value distributions were obtained from the highest free energy configurations along the lowest free energy paths between the unfolded and native states, as described in [6], except that additional terms representing hydrogen bonding and local sequence/structure preferences were included in the free energy function. The second β hairpin is favored by the computational model for protein G, because of an extensive hydrogen-bond network, and the first hairpin is favored by the model for protein L, because the second β turn has considerable torsional strain (three consecutive residues with positive phi angles).

symmetry, there is a definite hierarchy to structure formation. The characterization of the two transition states suggests that the lowest free energy route to the native state for this fold involves formation of one of the two β hairpins; however, the choice of hairpin is determined by factors beyond native state topology. Interestingly, with the addition of hydrogen bonding and sequence- and structure-dependent local free energies of ordering, the simple computational model described above [6] recapitulates the experimentally observed symmetry breaking (Figure 2).

The correspondence between the predicted and experimentally determined phi values suggests that the hairpin formed at the rate-limiting step is the one with the lowest free energy of formation. To test this hypothesis, computational protein design methods [18] have recently been used to specifically stabilize the first β hairpin of protein G, which, as noted above, is not formed in the transition state in the wild-type protein. A redesigned protein G variant with a more optimal backbone conformation and sequence in the first hairpin folds 100-fold faster than the wild-type protein. Subsequent mutational analysis shows that the first β hairpin, rather than the second β hairpin (as in the wild-type), is formed in the transition state in the redesigned protein (S Nauli, B Kuhlman, D Baker, unpublished data). Likewise, following stabilization by redesign of the second hairpin of protein L, which contains three consecutive residues with positive phi angles in the wild-type structure, and destabilization of the first hairpin, the second hairpin was found to be better formed in the folding transition-state ensemble than the first turn (D Kim, B Kuhlman, D Baker, unpublished data). These switches in folding mechanism highlight the differences local free energies of ordering can have when the overall topology has considerable symmetry.

AcP, Ada2h, U1A and S6

The folding transition states of four proteins with the ferredoxin-like fold (two helices packed against one side of a five-stranded β sheet) have been characterized. The folding transition states of Ada2h (activation domain of procarboxypeptidase) and AcP (acylphosphatase) are similar, despite the low sequence similarity (13%) between the two proteins and variations in the length of the secondary structural elements [19,20]. In both cases, the overall topology of the protein appears to be already specified in the transition state, but the second α helix and the inside strands of the β sheet with which it interacts appear to be more ordered than the rest of the polypeptide chain. The characterization of two other members of this structural family, however, revealed an alternative nucleus with preferential structure around helix 1: U1A nucleates in helix 1 and S6 nucleates in both helices [21]. The topology appears to allow several roughly equivalent folding pathways: the choice of the dominant pathway may be determined by the detailed packing and orientation of structural elements. Proteins with this fold also exhibit a pronounced movement of the transition state from 20% to 80% native (as judged by the burial of surface area) with increasing concentration of denaturant. Remarkably, given the variation in the transition-state structure, the folding rates of these proteins are highly correlated with the CO over an approximately 4000-fold range of folding rates. Furthermore, changing the CO can significantly change the folding rate: a circular permutant of U1A with CO lower than that of the wild-type protein folds considerably faster (M Oliveberg, personal communication).

SH3 domain fold

SH3 family

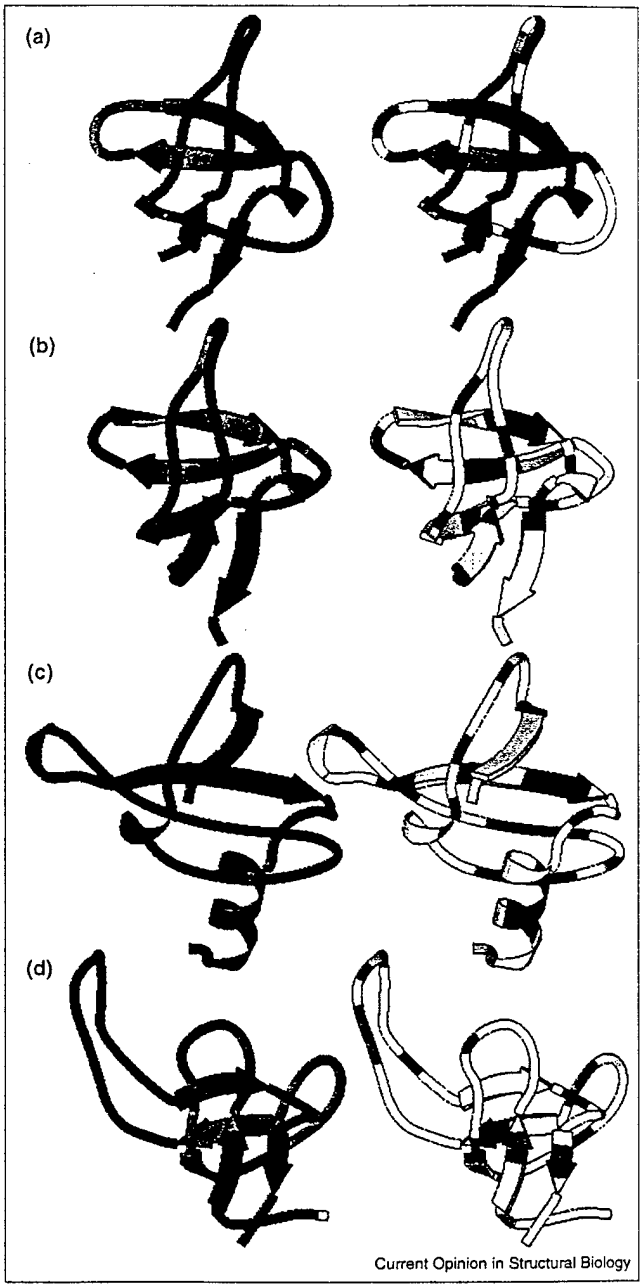
The homologous src and α-spectrin SH3 domains exhibit very similar transition states [3,22–24], despite the low

sequence identity (36%) (Figure 3a,b). Stabilizing mutations [23] and changes in pH [22] do not seem to affect the structure of the transition state of the α-spectrin SH3 domain. In the case of the src SH3 domain, stabilization of local structure by hairpin cross-linking and global stabilization by sodium sulfate do not alter the placement of the transition state along the reaction coordinate [25]. It appears, then, that SH3 domains allow quite large variations in sequence and experimental conditions with no change to the transition state, probably because there are no alternative structural elements that can be sufficiently stabilized to become folding nuclei. On the other hand, modifying the topology of the protein can significantly change the free energy landscape to favor alternative routes for folding. Circularization of the src SH3 domain causes the delocalization of structure in the transition state [25]. Circular permutation experiments on the α-spectrin SH3 domain also changed the transition state [26]. Connecting the wild-type termini with a small peptide linker and introducing a cut in the distal hairpin resulted in a shift in the structure of the transition state towards the n-src loop and the hairpin formed by the old termini; the former distal hairpin was completely disordered at the rate-limiting step. Therefore, shifts in transition-state structure can occur when formerly distant elements are covalently linked to reduce the entropic cost of their interaction. Drastic mutagenesis, which weakens the interaction energies throughout the protein, can also change the transition state. For example, a sequence-simplified mutant of the src SH3 domain made predominantly of five amino acids (isoleucine, lysine, glutamic acid, alanine and glycine) was found to have a more delocalized transition state (distal hairpin is not fully formed); the interactions stabilizing the wild-type SH3 transition state may not be strong enough in the simplified mutant to overcome the loss in entropy and residues from other parts of the protein may have to participate (Q Yi, D Baker, unpublished data).

SH3 structural analogs

The characterization of SH3 structural analogs has shown that transition-state structure is not always conserved in proteins with similar topologies. PsaE [27], a photosystem protein from cyanobacteria, has a large loop insertion at the distal hairpin (13 amino acids), making it entropically more costly to form stabilizing interactions. As a result, its transition state is more delocalized than that of the src SH3 domain, with well-ordered residues found in the distal hairpin, as well as in the N and C termini (P Bowers, D Baker, unpublished data) (Figure 3d). Sso7d, a DNA-binding protein from *Sulfolobus solfataricus* [28], has a significantly different transition state from that of the src and α-spectrin SH3 domains. The n-src loop and the C terminus (which is a helix in Sso7d, instead of a β strand) are the most structured in the transition state, whereas the distal hairpin is only weakly ordered (R Guerois, L Serrano, personal communication) (Figure 3c). This is in contrast to the src and α-spectrin SH3 transition states, in which the distal hairpin is completely ordered. In the SH3 domains and in Sso7d, the contiguous three-stranded sheet is formed but, in one case, the diverging turn interacts with it, whereas in the other case, it is the C-terminal helix. This difference may reflect variations in the free energies of forming the structural elements: in the SH3

Figure 3



Folding transition states of proteins with the SH3 fold: **(a)** src SH3 domain, **(b)** spectrin SH3 domain, **(c)** Sso7d and **(d)** PsaE. Left, predicted phi values (see legend to Figure 2); right, experimental phi values. The color scheme is described in the legend to Figure 2. The distal loop is favored over the n-src loop by the computational model for the src SH3 domain because it has more extensive hydrogen bonding, whereas the equivalent of the distal loop is disfavored by the model for Sso7d because it contains five glycine residues that are costly to order.

domains, the distal loop hairpin is well packed and the n-src loop is irregular, whereas in Sso7d, the opposite is the case — the equivalent of the distal hairpin contains five consecutive glycine residues (which are likely to be functionally important). With the inclusion of hydrogen bonding and sequence- and structure-dependent local free energies of ordering, the simple computational model described above [6] produces phi values very similar to those observed experimentally for the SH3 domains and Sso7d. Similar results were very recently published by Guerois and Serrano (R Guerois, L Serrano, unpublished data; see Now published).

In summary, folding transition-state structure is conserved more highly within the SH3 sequence superfamily than among SH3 analogs. The SH3 topology, then, although not as obviously symmetric as the protein L/protein G topology, still allows several alternative folding routes. The prevalence of one route over the other depends on the details of the structure. This may, in part, be due to the fact that functional constraints lead to the conservation within, but not between, superfamilies of portions of protein structures with unusual local features (the irregular n-src and RT loops in the SH3 domain, for example, are involved in proline-rich peptide binding) with higher free energies of formation. These features partially determine which of the pathways consistent with the native state topology is actually chosen.

The GCN4 and protein G experiments, together with the comparisons of transition-state structures in the AcP and SH3 families, suggest a picture in which several different 'pathways' with roughly equivalent free energy barriers can be consistent with the overall topology. Sequence changes can, by lowering or raising one barrier relative to another, produce significant changes in the transition-state ensemble without large changes in folding rate. Consistent with this picture, our most recent models of the folding process produce considerably more accurate predictions of folding transition-state structures when local free energies of ordering based on sequence-dependent backbone torsion angles and local hydrogen bonding terms are included. We anticipate considerable synergy between theory and experiment, and an important role for computational protein design methods in the further elucidation of the mechanisms of protein folding during the next few years.

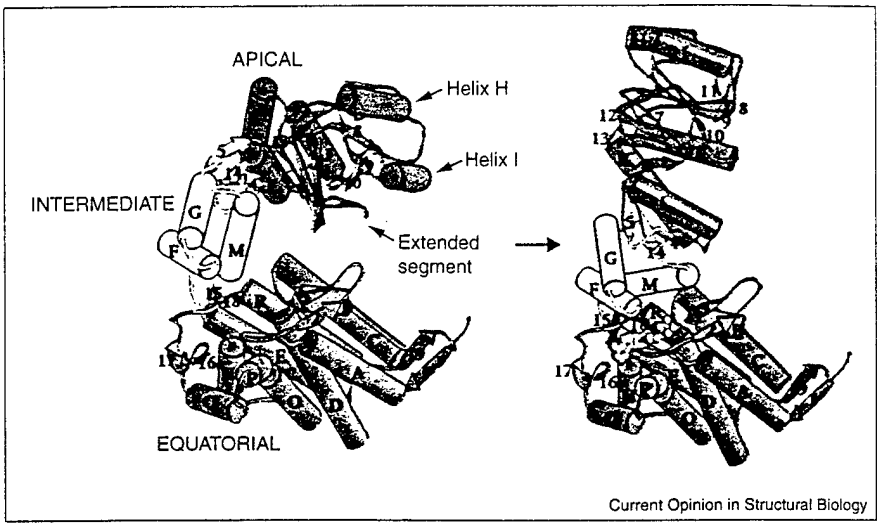
GroEL–GroES-assisted folding

How do the foregoing simple concepts apply to chaperone-assisted folding? In small proteins, the largest free energy barriers to folding involve the formation of particularly nonlocal portions of protein structures and regions with particularly unfavorable local energetics. It seems possible, therefore, that larger proteins containing such features may be particularly dependent on chaperones for suppressing alternative off-pathway misfolding/aggregation. Kinetic bottlenecks caused by unfavorable local structures or high contact order regions may tilt the kinetic competition between on- and off-pathway reactions in favor of the latter. It should be emphasized, however, that non-native interactions are likely to play a greater role in the folding of larger proteins simply because the increased size of the protein increases the probability of low free energy non-native interactions. Chaperones act on such non-native states in the first instance by binding the hydrophobic surfaces that are exposed, preventing these surfaces from 'wrongful interactions' that lead to multimolecular aggregation. Binding may, in some cases, be associated also with at least partial unfolding, as discussed below for GroEL. Release from the chaperones, in many cases driven by ATP binding (not hydrolysis), then allows the substrate polypeptide a chance to fold. Uniquely, in the case of the chaperonin ring class of chaperones, polypeptide is released into an encapsulated chamber where folding proceeds in isolation. In the case of the bacterial chaperonin, GroEL, this is mediated by ATP/GroES binding, which is associated with rigid-body movements of the GroEL intermediate and peptide-binding apical domains of the bound ring [29] (see Figure 4). The 60° elevation and 90° twisting of the apical domains act to remove the hydrophobic peptide-binding sites away from the central cavity, releasing polypeptide into this GroES-encapsulated space. Because the character of the wall of the cavity is switched from hydrophobic to hydrophilic as the result of the rigid-body movements, it may influence the released polypeptide to fold in this space because burial of exposed hydrophobic surfaces and exposure of hydrophilic surfaces, features of the native state, will be energetically favored.

Both cryo-EM reconstructions [30] and high-resolution crystal structures have resolved the rigid-body domain movements of the GroEL–GroES machinery itself during the reaction cycle [29,31] (see Figure 4). In addition, there are dynamic fluorescence and kinetic studies indicating, respectively, rapid release of bound polypeptide into the central cavity upon ATP/GroES binding ($t_{1/2}$ ~1 s) and productive folding inside the GroEL–GroES cavity [32–34]. However, the exact effects of the various states and transitions of the GroEL–GroES machinery during the reaction cycle on the conformation of polypeptide substrates are not well understood because, as ensembles of unstable non-native states, the substrates are much less accessible to structural study, particularly in the presence of the megadalton GroEL ring structure. Thus, our 'view' of what is happening to substrate proteins themselves during the GroEL–GroES reaction is poorly resolved. At this point, the study of stringent substrates, which are dependent on the complete system to reach their native form and are unable to productively fold without it, seems valuable for identifying and characterizing the full range of steps in the reaction that are critical to producing the native state. Nevertheless, there can also be value to studying nonstringent substrates, particularly those whose nonchaperoned folding is well described, because folding behavior can be compared in the presence and absence of

Figure 4

Rigid-body movements of a GroEL subunit attendant to ATP/GroES binding. Rigid-body rotations about the top and bottom of the intermediate domain redirect the peptide-binding surface of the apical domain, composed of helices H and I and an underlying extended segment, from a position facing the central cavity (lying to the right of the subunit) to a new position facing out of the page. The binding of peptides in the groove between helices H and I, through contacts with resident hydrophobic sidechains, has been observed (see text). Although the involvement of the extended segment of the apical domain in polypeptide binding has been indicated by mutational studies, a structural basis for such interaction remains undefined (adapted from [29]).



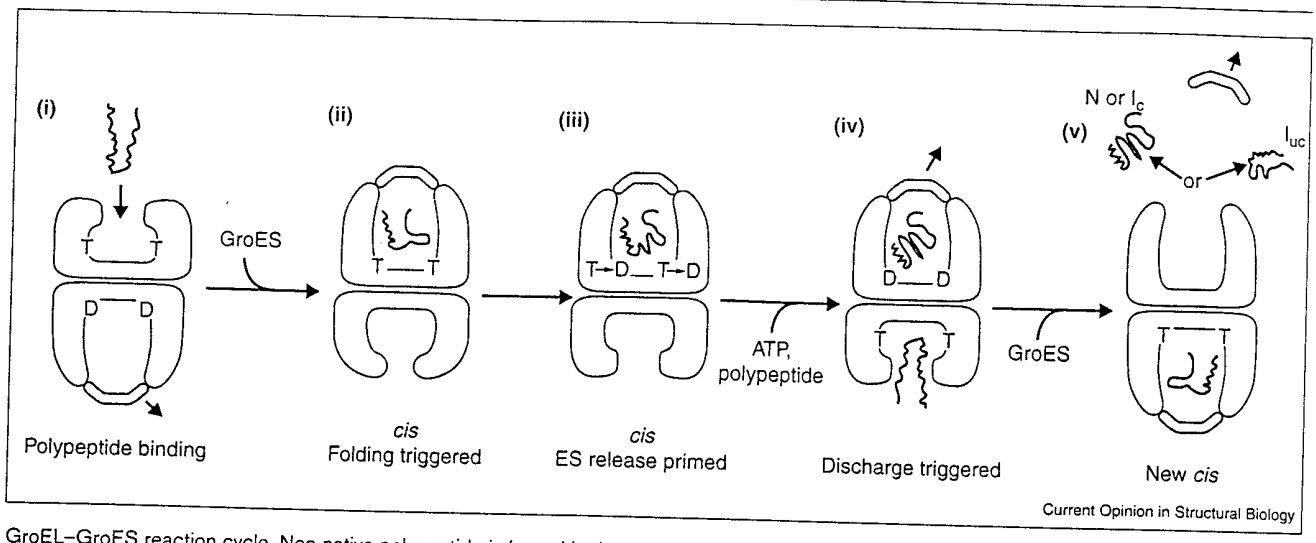
chaperonin. Even small peptides may, to some extent, simulate the behavior of a region of polypeptide chain, at least in binding to GroEL.

Binding to GroEL – potential unfolding action

There are definable points in the GroEL–GroES reaction cycle (Figure 5) at which major actions on polypeptide substrates have been considered likely to occur. One is the step of polypeptide binding to an open GroEL ring (which, under physiological conditions, would be the open ring of a GroEL–GroES–ADP asymmetric complex) [35] (see Figure 5). Binding may be associated with at least partial unfolding of a substrate protein, which is potentially a means for removing a non-native form from a kinetic trap. This could occur through either or both of two mechanisms, one catalytic, in which GroEL lowers the energy barriers between various non-native states, the other thermodynamic, in which GroEL preferentially binds less-folded states without affecting the transition states between the various conformations. The best evidence to date for a catalytic unfolding action associated with binding comes from a hydrogen-deuterium exchange experiment showing that GroEL in catalytic amounts can globally unfold the 6 kDa protein barnase [36]. Whether GroEL can exert similar effects on larger proteins, including those that form stable binary complexes with it, remains unclear. A number of exchange studies carried out with stable binary complexes of such proteins as α-lactalbumin [37], human dihydrofolate reductase [38,39] and Rubisco (ribulose-1,5-bisphosphate carboxylase-oxygenase) [40•] indicate that these proteins do not become globally exchanged while bound to GroEL, exhibiting modest levels of amide proton protection that are, in some cases, localized (but see, however, [41,42], which showed that cyclophilin and a chemically denatured β-lactamase, respectively, were completely exchanged while bound). In the case of Rubisco, it was possible to examine the protein both while in a metastable intermediate state in solution and after becoming bound to GroEL [40•]. In this case, a high degree of protection from exchange was observed for a small number of amide protons both in the metastable intermediate in solution and in the binary complex with GroEL. Thus, whatever the nature of this secondary structure(s), it appears to be resistant to the unfolding action associated with GroEL binding. Some proteins, however, may nevertheless be subject to catalyzed unfolding at a local level during the process of binding to GroEL.

The thermodynamic mechanism for unfolding in the presence of GroEL involves the greater affinity of GroEL for less-folded states among an ensemble of conformers that are in equilibrium with each other [43]. This would effectively shift the equilibrium by mass action toward the less-folded states. Perhaps the best evidence supporting an action of this sort comes from study of an RNase T1 mutant that populates two non-native states, one more structured than the other [44]. In the presence of GroEL, the less-folded state became more populated, without alteration of the microscopic rate constants between the two states, arguing for a thermodynamic effect (see also [42,45,46] for descriptions of such effects on β-lactamase, dihydrofolate reductase and barstar). Such partitioning between non-native states has yet to be demonstrated for stringent substrates, although the ability of GroEL to inhibit the production of off-pathway aggregates of malate dehydrogenase (MDH) has been kinetically modeled to such a mechanism. In the model, GroEL favors binding of MDH monomers and shifts an equilibrium of low-order aggregates of MDH toward this state [47]. Clearly, the ability to resolve different conformational states within an ensemble of substrate proteins, both unbound and GroEL-bound, using spectroscopic techniques, for example, will be necessary to better characterize the behavior of an open GroEL ring toward its substrates. Both catalytic and thermodynamic mechanisms could be

Figure 5



GroEL–GroES reaction cycle. Non-native polypeptide is bound in the open (*trans*) ring of an asymmetric GroEL–GroES–ADP (D) complex via hydrophobic interactions with the surrounding apical domains (panel i). Binding of ATP (T) and GroES to the same ring as the polypeptide produces large rigid-body movements in the subunits of the ring, elevating and twisting the hydrophobic binding surface away from the bound polypeptide, releasing it into the encapsulated and now hydrophilic *cis* chamber where folding commences (panel ii). After 8–10 s, ATP hydrolysis occurs in the seven subunits of the folding-active ring, relaxing the affinity of the ring for GroES and 'priming' it for release (panel iii). At the same time, *cis* hydrolysis produces an allosteric adjustment of the *trans* ring that allows rapid entry of ATP and non-native polypeptide (panel iv). The arrival of ATP triggers allosteric dissociation of the *cis* ligands (panel v); the binding of non-native polypeptide serves to accelerate the rate of this departure by 30–50-fold. Note that the polypeptide can be ejected in either a native form (N), a form committed to reaching the native state in the bulk solution (I_c) or an uncommitted non-native state (I_{uc}) that can be rebound by chaperonin. The relatively slow binding of GroES to the new ATP/polypeptide-bound ring orders the formation of the next folding-active GroEL–GroES complex (panel v). Thus, GroEL alternates rings back and forth as folding-active, expending the ATP of one ring to simultaneously initiate a new folding reaction, while dissociating the previous one from the opposite ring. As discussed in the text, polypeptide binding in an open GroEL ring (panels i and iv) may be associated with an action of unfolding. The step of ATP/GroES binding may also produce forced mechanical unfolding (panels ii and v).

operative, depending on the particular substrate and its position on the landscape. Finally, although the binding of substrate proteins is usually thought of as redirecting off-pathway states, there seems no reason to exclude that, in at least some cases, GroEL could recognize on-pathway intermediates, which could also receive kinetic assistance as a result of recruitment to the GroEL–GroES cavity.

Both catalytic and thermodynamic unfolding mechanisms could be enabled by the ability of the multiple surrounding GroEL apical domains to interact with a substrate protein. Such multivalent binding was recently indicated by an experiment with covalent GroEL rings bearing various numbers and arrangements of binding-proficient and binding-incompetent apical domains [48•]. A minimum of three consecutive proficient domains was required for efficient binding of a stringent substrate protein. In agreement, an accompanying experiment employing cysteine cross-linking between a bound substrate protein and a GroEL ring observed cross-links with multiple GroEL apical domains.

Translating binding action back to structure – what does GroEL recognize?

Ultimately, it would be desirable to translate the foregoing actions associated with chaperonin binding into structural terms. Lacking, however, any high-resolution information on the structure of a substrate protein bound to GroEL, we can only extrapolate from a variety of different types of experimental information, which, in the past year, has been derived from proteomic, biochemical, spectroscopic and crystallographic studies. At the level of binding to individual apical domains, a crystallographic study observed that a dodecamer peptide, selected for its high affinity for an isolated apical domain, associated with it as a β hairpin, both in a co-crystal with an isolated apical domain and in one with full occupancy of the apical domains of the GroEL tetradecamer [49•]. In these structures, one strand of the hairpin contacted the apical domain at a position between the two α helices (H and I) facing the central cavity (see Figure 4). A host of hydrophobic contacts were formed between tryptophan and phenylalanine residues in the peptide and hydrophobic sidechains in the two α helices; these helices had been previously implicated in polypeptide binding by a mutagenesis study [50] and by a previous crystallographic study of an apical domain [51]. In the latter study, similar topology and contacts were observed between an extended N-terminal tag segment of one monomer found lying in the groove between these two α helices in a neighboring monomer in the asymmetric unit. In the dodecamer study, it was additionally noted that, compared with the unoccupied isolated apical domain crystal structure, in which a

number of regions, including the channel-facing ones, were found to differ somewhat in positioning between monomers in the asymmetric unit, the conformations of the isolated domains with peptide bound became virtually identical. This suggests that there is a structural plasticity to the apical binding surface that accommodates the variety of substrates and that, upon contact with a particular substrate, optimizes contacts with it.

Lest it seem that only β strands can associate with the GroEL apical domain, two different NMR studies re-examined an N-terminal 13-residue peptide from the substrate rhodanese that is known to form an α helix in the intact native protein. This peptide had been observed through transfer NOE effects to adopt an α-helical structure upon association with intact GroEL [52]. In the first of the new studies, the same transfer NOE effects were observed when the peptide was incubated with an isolated GroEL apical domain, and chemical shift changes could be observed that localized to the same two cavity-facing α helices (H and I) [53]. In the second study, carried out with intact GroEL, D and D,L chiral forms of the same peptide were observed to bind as well as the original L form [54]. Whereas the D form could form a left-handed helix in TFE, the D,L form did not form α helix. This suggested that the hydrophobic content of the peptides was more critical to binding than adoption of a particular secondary structure. Two dodecameric α-helical peptides with the same composition were also compared, observing that one with hydrophobic sidechains clustered on one side of the predicted helix opposite hydrophilic sidechains (amphiphilic character) bound more strongly than another peptide interspersing hydrophobic sidechains with hydrophilic sidechains. This suggested that a contiguous hydrophobic surface is the feature in a substrate favoring its recruitment to GroEL. In a third study, a series of 14-residue peptides that exhibited α-helical character in solution was examined [55]. In this case also, those peptides with amphiphilic character were found to bind most strongly to GroEL, some with submicromolar affinity.

Thus, GroEL appears able to recognize both major secondary structural elements, so long as hydrophobic surface is presented. It remains curious, however, that, where examined, recognition appears to occur through the same two apical α helices without recognizable participation of an underlying extended segment (amino acids 199–209; see Figure 4) that also bears hydrophobic residues, mutation of which abolishes polypeptide binding. Thus, the question remains as to whether this segment participates directly in binding. Notably, the H and I α helices also form the major contacts with the GroES mobile loop (itself in an extended state), also through hydrophobic interactions, after elevation and twisting of the apical domains [29] (see Figure 4). Thus, binding through these two α helices may be an energetically favored mode, although polypeptide and GroES binding occur at two very different points in space.

Both major secondary structural elements figure together in a proteomic study identifying several dozen proteins from *Escherichia coli* that could be co-immunoprecipitated with anti-GroEL antiserum upon cell lysis in EDTA (to inhibit nucleotide-driven dissociation) [56]. Whether any of these are stringent substrates, that is, dependent on GroEL–GroES for proper folding, remains to be seen, but of this collective of bound species, where a structure of the native form was available, the topology favored was αβ, with two or more domains. Thus, it seems plausible that GroEL multivalently binds individual α and β units through exposed hydrophobic aspects that will be buried together in the native state. This potentially stabilizes the individual domains against inappropriate intermolecular or even intramolecular interactions until ATP/GroES-driven release directs an optimal chance for correct association within the molecule, while it is confined to the *cis* cavity. A direct illustration of such putative action comes from a study of the folding of four-disulfide hen lysozyme, composed of an α and β domain, in the presence of GroEL [57]. The open GroEL ring accelerated the rate of acquisition of the native state by 1.3-fold, without affecting the rate or mechanism of domain folding. Rather, GroEL accelerated the slower step of proper docking of the two domains, presumably by binding one or both individual domains and disfavoring or reversing non-native contacts.

ATP/GroES-driven release of GroEL-bound substrate into the central cavity – potential unfolding action

The action of ATP/GroES binding on polypeptide conformation, associated with release into the GroEL–GroES cavity, has been of major interest. An earlier study of the substrate Rubisco, examining its tryptophan fluorescence anisotropy, observed a rapid drop ($t_{1/2}$ ~1 s), followed by a slow rise correlating with production of the native state [32]. The nature of the fast phase had been a mystery, but an exchange experiment with tritium-labeled Rubisco has begun to address this [40•]. A metastable intermediate of this protein exhibited 12 highly protected amide tritiums both in solution and while bound to GroEL. When ATP and GroES were added, all but two of the tritiums were exchanged by 5 s, the earliest time examined. The elevation and twisting of the apical domains, driven by ATP/GroES binding to a polypeptide-bound ring, were proposed to produce a stretching of substrate between the apical domains before complete release into the cavity. Such a mechanism would couple the energy of ATP/GroES binding to a forced unfolding action. But the deprotection observed does not seem fully accountable only by a stretching action exerted on molecules becoming encapsulated in the *cis* ring. Consider the experimental observation that GroES binds randomly to either of the two GroEL rings of a Rubisco–GroEL binary complex to form two different asymmetric complexes: approximately 50% *cis* ternary complexes and approximately 50% *trans* ternary complexes, the latter with GroES on the ring opposite the polypeptide-bound one. Thus, one would expect

that, at a time (here, 5 s) less than that of a single turnover (~10 s), only about half of the tritiums should be deprotected, corresponding to those of the Rubisco molecules that had become encapsulated in *cis*. Yet nearly all were deprotected, suggesting that molecules in the *trans* ring must likewise have been deprotected. Previous studies have indicated that the *trans* ring of a *cis* complex in ATP has no significant affinity for non-native Rubisco [35], thus suggesting that any deprotection of Rubisco bound on that ring must be associated with its release into the bulk solution. Perhaps there is also a twisting action on that ring, attended by unfolding during release. Alternatively, simple release without unfolding may be sufficient to produce deprotection if, for example, the protection derives from association of the substrate with the GroEL cavity wall (either through direct hydrogen bond formation or via steric shielding of amide protons). Thus, more needs to be learned about whether forced unfolding is really occurring in this case, whether it is a general aspect of the chaperonin mechanism and whether substrate polypeptides bound in *trans* are somehow also affected. Furthermore, it remains to be demonstrated whether such an action is required for productive Rubisco folding.

In a further experiment, the kinetics of tritium exchange of the metastable Rubisco intermediate was examined in the presence of substoichiometric concentrations of GroEL–GroES. The observed rate of decay indicated that molecules whose tritiums had been exchanged were subsequently being released from *cis* complexes in non-native forms that competed with the remaining pool of still tritium-labeled Rubisco molecules for binding to GroEL [40•]. This reflects, as established by earlier studies, the occurrence of multiple rounds of binding and release of non-native polypeptide from GroEL during a productive folding reaction, underscoring the trial-and-error process of achieving the native state, as opposed to a process in which non-native forms remain at GroEL until productive folding occurs (see Figure 5). Indeed, in a stoichiometric reaction, only a few percent of Rubisco molecules reach native form in what corresponds to any given round of folding at chaperonin. Addition of 'trap' versions of GroEL, able to bind but not release non-native forms, rapidly halts a reaction, with non-native substrate physically accumulated at the trap (e.g. [58,59]). Such observations also reflect on the model for forced unfolding, indicating that, in and of itself, even if it occurs, such an action is not sufficient for producing the native state; otherwise, multiple rounds would not be required.

By contrast, when a stable, long-lived (>100 min) *cis* complex is formed between SR1, the single-ring version of GroEL, and GroES, it produces nearly 100% recovery of native Rubisco inside the *cis* cavity. This indicates a major role, if not a dominant one, for the encapsulated *cis* space in producing the native state (see also [60,61]). Furthermore, as suggested by kinetic studies with MDH, non-native molecules expelled into the bulk solution during a normal folding reaction with wild-type GroEL (where the lifetime of a *cis* complex is ~10 s) can form low-order aggregates on a short time-scale [47], in part explaining why such released forms fail to achieve the native state in the bulk solution. In contrast, MDH molecules held in a stable *cis* complex (inside SR1–GroES) are forestalled from such aggregation and are productively folded essentially quantitatively [32].

Productive folding in the GroEL–GroES cavity

Although features of the GroEL–GroES cavity that favor productive folding have been identified from crystallographic study, the path inside it that a protein takes to the native state is unknown. Does this chamber simulate an infinite dilution condition? Perhaps it can for smaller polypeptides, but the physical dimensions argue for close confinement of larger substrates like Rubisco. Experimentally, even in its native state, the smaller protein GFP appeared to be tumbling into the walls of this space, with a rotational correlation time of 42 ns, instead of the 12 ns observed in solution [33]. Perhaps such confinement presents limits to the conformational space that can be explored by non-native forms, limiting their folding trajectory. Clearly, a comparison of folding in this *cis* cavity with folding at infinite dilution would be instructive and might be possible using single-molecule techniques.

Conclusions

In sum, then, for both the folding of small two-state folding proteins and chaperonin action on larger ones, some basic outlines of mechanism are now available. Yet it seems likely that there will be still other basic mechanistic principles concerning these reactions that lie as yet unrecognized.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Plaxco KW, Simons KT, Baker D: **Contact order, transition state placement and the refolding rates of single domain proteins.** *J Mol Biol* 1998, **277**:985-994.
2. Plaxco KW, Simons KT, Ruczinski I, Baker D: **Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics.** *Biochemistry* 2000, **39**:11177-11183.
3. Riddle DS, Grantcharova VP, Santiago JV, Alm E, Ruczinski I, Baker D: **Experiment and theory highlight role of native state topology in SH3 folding.** *Nat Struct Biol* 1999, **6**:1016-1024.
4. Perl D, Welker C, Schindler T, Schroder K, Marahiel MA, Jaenicke R, Schmid FX: **Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins.** *Nat Struct Biol* 1998, **5**:229-235.
5. Plaxco KW, Guijarro JI, Morton CJ, Pitkeathly M, Campbell ID, Dobson CM: **The folding kinetics and thermodynamics of the Fyn-SH3 domain.** *Biochemistry* 1998, **37**:2529-2537.
6. Alm E, Baker D: **Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures.** *Proc Natl Acad Sci USA* 1999, **96**:11305-11310.
7. Debe DA, Goddard WA III: **First principles prediction of protein folding rates.** *J Mol Biol* 1999, **294**:619-625.
8. Galzitskaya OV, Finkelstein AV: **A theoretical search for folding/unfolding nuclei in three-dimensional protein structures.** *Proc Natl Acad Sci USA* 1999, **96**:11299-11304.

9. Muñoz V, Eaton WA: **A simple model for calculating the kinetics of protein folding from three-dimensional structures.** *Proc Natl Acad Sci USA* 1999, **96**:11311-11316.

10. Fersht AR, Matouschek A, Serrano L: **The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding.** *J Mol Biol* 1992, **224**:771-782.

11. Fersht AR: **Characterizing transition states in protein folding: an essential step in the puzzle.** *Curr Opin Struct Biol* 1995, **5**:79-84.

12. Moran LB, Schneider JP, Kentsis A, Reddy GA, Sosnick TR: **Transition state heterogeneity in GCN4 coiled coil folding studied by using multisite mutations and crosslinking.** *Proc Natl Acad Sci USA* 1999, **96**:10699-10704.

13. Zitzewitz JA, Ibarra-Molero B, Fishel DR, Terry KL, Matthews CR: **Preformed secondary structure drives the association reaction of GCN4-p1, a model coiled-coil system.** *J Mol Biol* 2000, **296**:1105-1116.

14. Burton RE, Huang GS, Daugherty MA, Calderone TL, Oas TG: **The energy landscape of a fast-folding protein mapped by Ala→Gly substitutions.** *Nat Struct Biol* 1997, **4**:305-310.

15. Myers JK, Oas TG: **Contribution of a buried hydrogen bond to lambda repressor folding kinetics.** *Biochemistry* 1999, **38**:6761-6768.

16. Kim DE, Fisher C, Baker D: **A breakdown of symmetry in the folding transition state of protein L.** *J Mol Biol* 2000, **298**:971-984.

17. McCallister EL, Alm E, Baker D: **Critical role of beta-hairpin formation in protein G folding.** *Nat Struct Biol* 2000, **7**:669-673.

18. Kuhlman B, Baker D: **Native protein sequences are close to optimal for their structures.** *Proc Natl Acad Sci USA* 2000, **97**:10383-10388.

19. Chiti F, Taddei N, White PM, Bucciantini M, Magherini F, Stefani M, Dobson CM: **Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding.** *Nat Struct Biol* 1999, **6**:1005-1009.

20. Taddei N, Chiti F, Fiaschi T, Bucciantini M, Capanni C, Stefani M, Serrano L, Dobson CM, Ramponi G: **Stabilisation of alpha-helices by site-directed mutagenesis reveals the importance of secondary structure in the transition state for acylphosphatase folding.** *J Mol Biol* 2000, **300**:633-647.

21. Ternstrom T, Mayor U, Akke M, Oliveberg M: **From snapshot to movie: phi analysis of protein folding transition states taken one step further.** *Proc Natl Acad Sci USA* 1999, **96**:14854-14859.

22. Martinez JC, Serrano L: **The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved.** *Nat Struct Biol* 1999, **6**:1010-1016.

23. Martinez JC, Pisabarro MT, Serrano L: **Obligatory steps in protein folding and the conformational diversity of the transition state.** *Nat Struct Biol* 1998, **5**:721-729.

24. Grantcharova VP, Riddle DS, Baker D: **Long-range order in the src SH3 folding transition state.** *Proc Natl Acad Sci USA* 2000, **97**:7084-7089.

25. Grantcharova VP, Baker D: **Circularization changes the folding transition state of the src SH3 domain.** *J Mol Biol* 2001, in press.

26. Viguera AR, Jimenez MA, Rico M, Serrano L: **Conformational analysis of peptides corresponding to beta-hairpins and a beta-sheet that represent the entire sequence of the alpha-spectrin SH3 domain.** *J Mol Biol* 1996, **255**:507-521.

27. Falzone CJ, Kao YH, Zhao J, Bryant DA, Lecomte JT: **Three-dimensional solution structure of PsaE from the cyanobacterium *Synechococcus* sp. strain PCC 7002, a photosystem I protein that shows structural homology with SH3 domains.** *Biochemistry* 1994, **33**:6052-6062.

28. Baumann H, Knapp S, Lundback T, Ladenstein R, Hard T: **Solution structure and DNA-binding properties of a thermostable protein from the archaeon *Sulfolobus solfataricus*.** *Nat Struct Biol* 1994, **1**:808-819.

29. Xu Z, Horwich AL, Sigler PB: **The crystal structure of the asymmetric GroEL-GroES-$(ADP)_7$ chaperonin complex.** *Nature* 1997, **388**:741-750.

30. Roseman AM, Chen S, White H, Braig K, Saibil HR: **The chaperonin ATPase cycle: mechanism of allosteric switching and movements of substrate-binding domains in GroEL.** *Cell* 1996, **87**:241-251.

31. Braig K, Otwinowski Z, Hegde R, Boisvert DC, Joachimiak A, Horwich AL, Sigler PB: **The crystal structure of the bacterial chaperonin GroEL at 2.8 Å.** *Nature* 1994, **371**:578-586.

32. Rye HS, Burston SG, Fenton WA, Beechem JM, Xu Z, Sigler PB, Horwich AL: **Distinct actions of cis and trans ATP within the double ring of the chaperonin GroEL.** *Nature* 1997, **388**:792-798.

33. Weissman JS, Rye HS, Fenton WA, Beechem JM, Horwich AL: **Characterization of the active intermediate of a GroEL-GroES-mediated protein folding reaction.** *Cell* 1996, **84**:481-490.

34. Ranson NA, Burston SG, Clarke AR: **Binding, encapsulation and ejection: substrate dynamics during a chaperonin-assisted folding reaction.** *J Mol Biol* 1997, **266**:656-664.

35. Rye HS, Roseman AM, Chen S, Furtak K, Fenton WA, Saibil HR, Horwich AL: **GroEL-GroES cycling: ATP and nonnative polypeptide direct alternation of folding-active rings.** *Cell* 1999, **97**:325-338.

36. Zahn R, Perrett S, Stenberg G, Fersht AR: **Catalysis of amide proton exchange by the molecular chaperones GroEL and SecB.** *Science* 1996, **271**:642-645.

37. Robinson CV, Gross M, Eyles SJ, Ewbank JJ, Mayhew M, Hartl FU, Dobson CM, Radford SE: **Conformation of GroEL-bound alpha-lactalbumin probed by mass spectrometry.** *Nature* 1994, **372**:646-651.

38. Gross M, Robinson CV, Mayhew M, Hartl FU, Radford SE: **Significant hydrogen exchange protection in GroEL-bound DHFR is maintained during iterative rounds of substrate cycling.** *Protein Sci* 1996, **5**:2506-2513.

39. Goldberg MS, Zhang J, Sondek S, Matthews CR, Fox RO, Horwich AL: **Native-like structure of a protein-folding intermediate bound to the chaperonin GroEL.** *Proc Natl Acad Sci USA* 1997, **94**:1080-1085.

40. • Shtilerman M, Lorimer GH, Englander SW: **Chaperonin function: folding by forced unfolding.** *Science* 1999, **284**:822-825.
Evidence from tritium exchange that Rubisco is 'stretched on the rack' in the act of ATP/GroES binding to a Rubisco-bound GroEL ring.

41. Zahn R, Spitzfaden C, Ottiger M, Wuthrich K, Pluckthun A: **Destabilization of the complete protein secondary structure on binding to the chaperone GroEL.** *Nature* 1994, **368**:261-265.

42. Gervasoni P, Gehrig P, Pluckthun A: **Two conformational states of beta-lactamase bound to GroEL: a biophysical characterization.** *J Mol Biol* 1998, **275**:663-675.

43. Zahn R, Pluckthun A: **Thermodynamic partitioning model for hydrophobic binding of polypeptides by GroEL. II. GroEL recognizes thermally unfolded mature beta-lactamase.** *J Mol Biol* 1994, **242**:165-174.

44. Walter S, Lorimer GH, Schmid FX: **A thermodynamic coupling mechanism for GroEL-mediated unfolding.** *Proc Natl Acad Sci USA* 1996, **93**:9425-9430.

45. Clark AC, Frieden C: **The chaperonin GroEL binds to late-folding non-native conformations present in native *Escherichia coli* and murine dihydrofolate reductases.** *J Mol Biol* 1999, **285**:1777-1788.

46. Bhutani N, Udgaonkar JB: **A thermodynamic coupling mechanism can explain the GroEL-mediated acceleration of the folding of barstar.** *J Mol Biol* 2000, **297**:1037-1044.

47. Ranson NA, Dunster NJ, Burston SG, Clarke AR: **Chaperonins can catalyse the reversal of early aggregation steps when a protein misfolds.** *J Mol Biol* 1995, **250**:581-586.

48. • Farr GW, Furtak K, Rowland MB, Ranson NA, Saibil HR, Kirchhausen T, Horwich AL: **Multivalent binding of nonnative substrate proteins by the chaperonin GroEL.** *Cell* 2000, **100**:561-573.
Functional and physical evidence that polypeptide substrates are bound by multiple consecutive apical domains of an open GroEL ring.

49. • Chen L, Sigler PB: **The crystal structure of a GroEL/peptide complex: plasticity as a basis for substrate diversity.** *Cell* 1999, **99**:757-768.
Exogenously bound dodecamer peptide in a binary complex with the GroEL apical domains forms hydrophobic contacts and a few hydrogen bonds.

50. Fenton WA, Kashi Y, Furtak K, Horwich AL: **Residues in chaperonin GroEL required for polypeptide binding and release.** *Nature* 1994, **371**:614-619.

51. Buckle AM, Zahn R, Fersht AR: **A structural model for GroEL-polypeptide recognition.** *Proc Natl Acad Sci USA* 1997, **94**:3571-3575.

52. Landry SJ, Gierasch LM: **The chaperonin GroEL binds a polypeptide in an alpha-helical conformation.** *Biochemistry* 1991, **30**:7359-7362.

53. Kobayashi N, Freund SM, Chatellier J, Zahn R, Fersht AR: **NMR analysis of the binding of a rhodanese peptide to a minichaperone in solution.** *J Mol Biol* 1999, **292**:181-190.

54. Wang Z, Feng H, Landry SJ, Maxwell J, Gierasch LM: **Basis of substrate binding by the chaperonin GroEL.** *Biochemistry* 1999, **38**:12537-12546.

55. Preuss M, Hutchinson JP, Miller AD: **Secondary structure forming propensity coupled with amphiphilicity is an optimal motif in a peptide or protein for association with chaperonin 60 (GroEL).** *Biochemistry* 1999, **38**:10272-10286.

56. Houry WA, Frishman D, Eckerskorn C, Lottspeich F, Hartl FU: **Identification of *in vivo* substrates of the chaperonin GroEL.** *Nature* 1999, **402**:147-154.

57. Coyle JE, Texter FL, Ashcroft AE, Masselos D, Robinson CV, Radford SE: **GroEL accelerates the refolding of hen lysozyme without changing its folding mechanism.** *Nat Struct Biol* 1999, **6**:683-690.

58. Weissman JS, Kashi Y, Fenton WA, Horwich AL: **GroEL-mediated protein folding proceeds by multiple rounds of binding and release of nonnative forms.** *Cell* 1994, **78**:693-702.

59. Todd MJ, Viitanen PV, Lorimer GH: **Dynamics of the chaperonin ATPase cycle: implications for facilitated protein folding.** *Science* 1994, **265**:659-666.

60. Beissinger M, Rutkat K, Buchner J: **Catalysis, commitment and encapsulation during GroE-mediated folding.** *J Mol Biol* 1999, **289**:1075-1092.

61. Grallert H, Buchner J: **Analysis of GroE-assisted folding under nonpermissive conditions.** *J Biol Chem* 1999, **274**:20171-20177.

62. Wittung-Stafshede P, Lee JC, Winkler JR, Gray HB: **Cytochrome b562 folding triggered by electron transfer: approaching the speed limit for formation of a four-helix-bundle protein.** *Proc Natl Acad Sci USA* 1999, **96**:6587-6590.

63. Ghaemmaghami S, Word JM, Burton RE, Richardson JS, Oas TG: **Folding kinetics of a fluorescent variant of monomeric lambda repressor.** *Biochemistry* 1998, **37**:9179-9185.

64. Spector S, Raleigh DP: **Submillisecond folding of the peripheral subunit-binding domain.** *J Mol Biol* 1999, **293**:763-768.

65. Mines GA, Pascher T, Lee SC, Winkler JR, Gray HB: **Cytochrome c folding triggered by electron transfer.** *Chem Biol* 1996, **3**:491-497.

66. Ferguson N, Capaldi AP, James R, Kleanthous C, Radford SE: **Rapid folding with and without populated intermediates in the homologous four-helix proteins Im7 and Im9.** *J Mol Biol* 1999, **286**:1597-1608.

67. Kragelund BB, Osmark P, Neergaard TB, Schiodt J, Kristiansen K, Knudsen J, Poulsen FM: **The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP.** *Nat Struct Biol* 1999, **6**:594-601.

68. Choe SE, Matsudaira PT, Osterhout J, Wagner G, Shakhnovich EI: **Folding kinetics of villin 14T, a protein domain with a central beta-sheet and two hydrophobic cores.** *Biochemistry* 1998, **37**:14508-14518.

69. Kuhlman B, Luisi DL, Evans PA, Raleigh DP: **Global analysis of the effects of temperature and denaturant on the folding and unfolding kinetics of the N-terminal domain of the protein L9.** *J Mol Biol* 1998, **284**:1661-1670.

70. Khorasanizadeh S, Peters ID, Butt TR, Roder H: **Folding and stability of a tryptophan-containing mutant of ubiquitin.** *Biochemistry* 1993, **32**:7054-7063.

71. Itzhaki LS, Otzen DE, Fersht AR: **The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding.** *J Mol Biol* 1995, **254**:260-288.

72. Silow M, Oliveberg M: **High-energy channeling in protein folding.** *Biochemistry* 1997, **36**:7633-7637.

73. Villegas V, Martinez JC, Aviles FX, Serrano L: **Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain.** *J Mol Biol* 1998, **283**:1027-1036.

74. Smith CK, Bu Z, Anderson KS, Sturtevant JM, Engelman DM, Regan L: **Surface point mutations that significantly alter the structure and stability of a protein's denatured state.** *Protein Sci* 1996, **5**:2009-2019.

75. Scalley ML, Yi Q, Gu H, McCormack A, Yates JR III, Baker D: **Kinetics of folding of the IgG binding domain of peptostreptococcal protein L.** *Biochemistry* 1997, **36**:3373-3382.

76. Main ER, Fulton KF, Jackson SE: **Folding pathway of FKBP12 and characterisation of the transition state.** *J Mol Biol* 1999, **291**:429-444.

77. van Nuland NA, Meijberg W, Warner J, Forge V, Scheek RM, Robillard GT, Dobson CM: **Slow cooperative folding of a small globular protein HPr.** *Biochemistry* 1998, **37**:622-637.

78. Aronsson G, Brorsson AC, Sahlman L, Jonsson BH: **Remarkably slow folding of a small protein.** *FEBS Lett* 1997, **411**:359-364.

79. van Nuland NA, Chiti F, Taddei N, Raugei G, Ramponi G, Dobson CM: **Slow folding of muscle acylphosphatase in the absence of intermediates.** *J Mol Biol* 1998, **283**:883-891.

80. Clarke J, Cota E, Fowler SB, Hamill SJ: **Folding studies of immunoglobulin-like beta-sandwich proteins suggest that they share a common folding pathway.** *Structure* 1999, **7**:1145-1153.

81. Viguera AR, Serrano L, Wilmanns M: **Different folding transition states may result in the same native structure.** *Nat Struct Biol* 1996, **3**:874-880.

82. Hamill SJ, Steward A, Clarke J: **The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology.** *J Mol Biol* 2000, **297**:165-178.

83. Otzen DE, Fersht AR: **Folding of circular and permuted chymotrypsin inhibitor 2: retention of the folding nucleus.** *Biochemistry* 1998, **37**:8139-8146.

84. Fulton KF, Main ER, Daggett V, Jackson SE: **Mapping the interactions present in the transition state for unfolding/folding of FKBP12.** *J Mol Biol* 1999, **291**:445-461.

85. López-Hernández E, Serrano L: **Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein, CI-2.** *Fold Des* 1995, **1**:43-55.

86. Schymkowitz JW, Rousseau F, Irvine LR, Itzhaki LS: **The folding pathway of the cell-cycle regulatory protein p13suc1: clues for the mechanism of domain swapping.** *Structure* 2000, **8**:89-100.

87. Milla ME, Brown BM, Waldburger CD, Sauer RT: **P22 arc repressor: transition state properties inferred from mutational effects on the rates of protein unfolding and refolding.** *Biochemistry* 1995, **34**:13914-13919.

88. Kraulis PJ: **MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures.** *J Appl Crystallogr* 1991, **24**:946-950.

Now published

The work referred to in the text as (R Guerois, L Serrano, unpublished data) has now been published:

89. Guerois R, Serrano L: **The SH3-fold family: experimental evidence and prediction of variations in the folding pathways.** *J Mol Biol* 2000, **304**:987-982.